AD__________________

GRANT NUMBER: DAMD17-94-J-4461

TITLE: **Methodology for Case-Control Studies of Breast Cancer**

PRINCIPAL INVESTIGATOR: **Donna Jean Brogan, Ph.D.**

CONTRACTING ORGANIZATION: Rollins School of Public Health
Emory University
Atlanta, GA 30322

REPORT DATE: October 1996

TYPE OF REPORT: Final

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, MD 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1996	Final (1 Sep 94 - 30 Sep 96)

4. TITLE AND SUBTITLE
Methodology for Case-Control Studies of Breast Cancer

5. FUNDING NUMBERS
DAMD17-94-J-4461

6. AUTHOR(S)

Donna Jean Brogan, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Rollins School of Public Health
Emory University
Atlanta, GA 30322

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, MD 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT

This project investigated whether the use of random digit dialing (RDD) sampling for obtaining a control group in a breast cancer case-control study incurred any substantial biases, compared to the use of area probability sampling. Data from a case-control study on metropolitan Atlanta women aged 20-54 years were analyzed, with two independently obtained control groups (area and RDD). The two sample control groups both agreed with sample Census data in making inference to the larger population. In unweighted analyses the two control groups differed somewhat on race, with a larger percentage of Blacks in the area sample. However, after adjustment for age and race, the two control groups were remarkably similar on 41 variables related to breast cancer risk and general health status. Two notable features of RDD sampling, compared to area sampling, were a significantly smaller enumeration or screening response rate and a significant under-identification of households which contained any women aged 20-54 years. Although this study provides no compelling evidence for caution regarding RDD sampling, an awareness of its generally higher nonresponse at all stages of public contact is recommended.

14. SUBJECT TERMS
Breast Cancer

15. NUMBER OF PAGES
58

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the US Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

DJB ✓ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products of services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Donna Jean Brogan 7-28-97

PI - Signature Date

TABLE OF CONTENTS

1. INTRODUCTION

1. 1 Background

The case-control study is the primary method of identification of potential risk factors in breast cancer epidemiology. Selection of a control group for breast cancer case-control studies has a long developmental history, beginning with hospital controls and followed by neighborhood controls. More recently population- based probability samples have been selected from the community, county, or state from which the cases derive in order to reduce the well-described potential biases that can result from the use of nonpopulation-based controls (Schlesselman, 1982; Rothman, 1986). Two common methods for selecting a population-based control group are area sampling and random digit dialing (RDD) sampling.

In area sampling a probability sample of geographic areas (such as blocks or block groups) is selected from the population, followed by a probability sample of housing units (HUs) from each selected geographic area, followed by a probability sample of one or more eligible persons within each selected HU (Kish, 1965; Cochran, 1977). In RDD sampling typically a probability sample of "telephone banks" (clusters of telephone numbers) is selected, followed by a probability sample of residential telephone numbers within the bank, followed by a probability sample of one or more eligible persons within the HU identified by the selected telephone number (Waksberg, 1978; Hartge, et al., 1984). Both techniques involve clustering (of telephone numbers or of HUs), and an equal probability sample of persons is not possible if only one person is selected from each HU. More recent RDD techniques use list-assisted sampling whereby the sampled telephone numbers are not clustered, often in combination with telephone numbers stratified by likelihood of being residential and some of the strata not even sampled (Casady and Lepkowski, 1991; Potter et al, 1991).

Both RDD and area sampling are expected to be more expensive than sampling techniques which use hospital or neighborhood controls. RDD sampling tends to be less expensive than area sampling in many instances (Groves and Kahn, 1979; Groves, 1989), although much of the research in this area has compared the cost of telephone interviews conducted in RDD samples with personal, or face-to-face, interviews in area samples (Groves, 1989). In addition, it is easier to supervise and ensure the security of people working at a telephone center (RDD sampling) than of those working in the field counting and listing HUs (area sampling). Thus, RDD sampling has become a fairly standard method for many public health studies, including the selection of control samples in case-control studies of cancer (Harlow and Hartge, 1983; Hartge, et al, 1984; Blot, et al., 1988; Wingo, et al., 1988).

Several potential sources of bias have been attributed to RDD. First, it has been suggested that RDD selected respondents differ from RDD selected non-respondents to a greater extent than for other sampling techniques (Groves and Lyberg, 1988), particularly

among minorities, lower education groups, and the elderly (Aquilino and Losciuto, 1990; Olsen and Mandel, 1988; Groves, et al., 1987; Freeman, et al., 1982). However, these studies did not attempt to separate the effects of different interviewing modalities (telephone vs. personal) from the effects arising from different sampling techniques (telephone vs. area).

Since RDD sampling frames do not cover persons in the population who live in HUs without a residential telephone, this may be a second source of bias. Estimates of telephone noncoverage in the United States for 1986 range from 4 to 21 percent across states (Trewin and Lee, 1988). The percentage of United States households without telephones has decreased over time from approximately 20 percent in 1963 to 10 percent by the early 1970s, and has stabilized at around 7 percent since the early 1980s (Thornberry and Massey, 1988). Telephone noncoverage appears to be higher in rural areas and in inner cities, and is higher for single person and large households, low income groups, and households containing unemployed persons and young heads of households (Trewin and Lee, 1988). Noncoverage rates also vary by race, ranging from 6 percent for whites to 16 percent for African-Americans and 19 percent for Hispanics (Thornberry and Massey, 1988).

Households in the southern United States are less likely to have telephones than those in other geographic areas (Thornberry and Massey, 1988). Between 1985 and 1986, noncoverage in southern Metropolitan Statistical Areas ranged from 9.6 percent for central city households to 7.7 percent for non-central city households (Thornberry and Massey, 1988). Telephone noncoverage in this region is 8.8 percent for whites, 19.9 percent for African-Americans, and 10.8 percent for other races (Thornberry and Massey, 1988). Other sociodemographic variables related to telephone noncoverage in the southern United States include highest educational attainment of less than 12 years for a responsible adult family member (24 percent), family income under $ 15,000 (19 percent or more), unemployed adult family member (23 percent), and marital status of separated (24.8 percent), married with spouse not in the household (20.5 percent), or divorced (13.5 percent) (Thornberry and Massey, 1988).

A few studies have examined differences in health-related characteristics between nontelephone and telephone households. Thornberry and Massey (1988), using National Health Interview Survey data from 1963 through 1986, estimated that the nontelephone population is more likely to have chronic health conditions and lower rates of utilization of health services, except for hospitalizations, relative to the population with telephone coverage. The population of persons under age 65 living in nontelephone households is less likely, relative to those under 65 in telephone households, to have private health insurance (35 vs 80 percent) or health insurance coverage of any type (61 vs 87 percent). Persons residing in nontelephone households are more likely to smoke cigarettes (50 vs 29 percent) and less likely to exercise regularly (32 vs 41 percent). Olson, et al. (1992) compared data collected using a private population census conducted in Otsego County, New York to responses from personal interviews of participants contacted via RDD. The

RDD respondents were more likely to have had their cholesterol level tested within the past two years, and among females, to have ever had a mammogram.

Because telephone ownership varies by geographic area, socioeconomic status, and race, and because a number of health-related factors may be associated with residence in a nontelephone household, the use of RDD sampling may have more potential for bias in some surveys than in others. However, even if the RDD sample is biased to some extent, this bias may have negligible impact on important analyses such as estimating the prevalence of breast cancer risk factors or estimating the odds ratios for identified risk factors for breast cancer. Although these concerns have been voiced almost since the inception of RDD sampling, to our knowledge no studies have clearly addressed the potential bias of RDD sampling in case-control studies of cancer (Wingo, et al., 1988; Longnecker, 1989).

Hence, even though RDD is commonly used for selecting a population-based control group for cancer case-control studies, concerns persist about its potential bias. This methodological and empirical study provides unique data with which to investigate potential biases of RDD samples within the context of a case-control study which investigated breast cancer risk factors in women aged 20-54 years.

WISH (Women's Interview Study of Health) was a multicenter breast cancer case-control study funded by the Environmental Epidemiology Branch of NCI during 1989-93 (Brinton et al, 1995). The study was restricted to younger women aged 20-54 years with special emphasis on risk factors such as exogenous hormones, diet, anthropometry, alcohol consumption and medical factors. Investigators at the Atlanta site collected interview data, under contract to NCI, on breast cancer cases (n = 777) and **two** age frequency-matched control groups: RDD controls (n = 652) selected by the Westat organization and area (probability sampling) controls (n = 640) selected by investigators at Rollins School of Publich Health at Emory University. Procedures for selecting the two statistically independent control groups were designed to be as comparable as possible through collaboration of Emory investigators with Westat and NCI personnel. All three samples—cases, RDD controls and area controls—were statistically independent. Although different teams of survey staff performed the RDD and area sampling, the same team of Atlanta interviewers conducted face-to-face interviews for the Atlanta cases and the two Atlanta control samples. Identical survey instruments and anthropometric measurements were used for all respondents. Hence, a comparison of the two control samples, RDD and area, is a comparison of the two sampling methodologies, unconfounded by type of interview, interviewing staff, and geographic location.

1.2. Specific Aims and Significance

The first specific aim was to compare random digit dialing (RDD) sampling with area sampling for the purpose of obtaining a population-based control group for a breast cancer case-control study. This specific aim was evaluated by assessing whether these two standard probability sampling methodologies yielded control groups which

1) are fairly equivalent to each other on enumeration response rates, interview response rates, and survey response rates,

2) are fairly equivalent to the 1990 census and to each other on demographic and social characteristics common to the census and the surveys,

3) are fairly equivalent on prevalences of risk factors for breast cancer,

4) yield fairly equivalent odds ratios for the salient breast cancer risk factors.

The second specific aim was for the investigator, Dr. Brogan, to enhance her skills and experience in the conduct of breast cancer research, particularly of an epidemiological nature, so that she could change the focus of the next phase of her career from biostatistics administration to research in two areas: breast cancer as a new emphasis and complex sample surveys as a continuing and long-standing interest.

These two specific aims were accomplished by Dr. Brogan spending a sabbatical year at the National Cancer Institute (NCI/NIH in Rockville, MD) under the sponsorship of Dr. Louise Brinton, with followup analytical effort at Emory and continuing collaboration with NCI colleagues during a no-cost extension of the Army grant into a second year. The Army grant paid approximately one-third of Dr. Brogan's salary during these two years, and NCI paid about 10% of her salary for 9 of the 12 months she spent at NCI.

2. BODY-METHODS

2.1 Overview of WISH Study Methods

In 1989 NCI funded a contract to Atlanta, Seattle and New Jersey to conduct a multi-site case-control study to assess breast cancer risk factors among younger women aged 20-44 years, with particular emphasis on early and long term use of oral contraceptives, alcohol consumption, anthropometry, and dietary intake as an adult and an adolescent. Because of longstanding concern about the potential biases in RDD methodology, NCI also funded procurement of a second or "alternate" control group, using area probability sampling, at the Atlanta site only. In addition, only the Atlanta site expanded the study's age range to 20-54 years.

Each site identified and selected its breast cancer cases to be interviewed. Westat performed RDD sampling and identified age frequency-matched RDD controls to be interviewed at all three sites, where the age-matching was done based on the anticipated age distribution of the cases. Emory investigators conducted area probability sampling to identify age-matched area controls to be interviewed at the Atlanta site. Emory and Westat investigators collaborated from the inception of the WISH study to make their

sampling procedures as comparable as possible so as to maximize the benefit from the methodological study.

All interviewing (cases, RDD controls, and area controls) was performed face-to-face by female interviewers, most often in the subject's home. The interview instrument contained the following sections: background (demographic) information, pregnancy history, menstruation and menopause history, contraceptive history, hormone medication history, medical history, developmental history and physical activity, adolescent diet, alcohol consumption, smoking history, occupational history, family history, and lifestyle and opinion. After the interview the woman self-completed a 19 page food questionnaire covering the last 12 months and answered questions about changes in eating habits during the past 10-15 years.

Atlanta RDD controls were assigned by Westat for interview between May, 1990 and May, 1992. Emory investigators identified area controls for interview between Sept., 1990 and March, 1992. Interviewing of RDD controls occurred between June, 1990 and October, 1992, with the time period for interview of area controls ranging from Sept. 1990 through Oct. 1992. All three sites extended data collection for eight months beyond the original contract termination date of October, 1992. However, the extension at the Atlanta site did not apply to selection and interview of additional area controls, and the present research includes only cases and controls selected during the initial two years of the contract.

The sample size of interviewed women for the analyses conducted under the Army grant is as follows: 640 area controls, 652 RDD controls and 777 cases.

2.2 RDD Sampling Methodology for the Atlanta WISH Site

Westat used the Waksberg (1978) RDD method to select a probability sample of about 900 women aged 20-54 from Fulton, DeKalb and Cobb counties in metropolitan Atlanta. There was no attempt to match the RDD controls (or the area controls) to the cases on race, no stratification was used, and only one area code (404) was used for the RDD sample (the only area code in the three counties at the time of fielding the WISH study).

Details of the well known Mitofsky-Waksberg procedure are not given here. In general the sampling procedure selects an equal probability sample of telephone banks (a bank is a cluster of 100 telephone numbers differing only in their last two digits), retains a given bank in the sample proportional to the number of residential telephone numbers in the bank, and then samples a fixed number of residential telephone numbers within retained banks. This procedure yields an approximately equal probability sample of households with a residential telephone. (Since some HUs have two or more residential telephone numbers, the equal probability sample of telephone numbers is an "approximately" equal probability sample of HUs.)

Once a residential telephone number was reached, enumeration or screening of household members was conducted over the telephone by determining if the household had any female members aged 20-54 years. Sampling within the household was not conducted at the time of telephone enumeration but subsequently by Westat so that the age frequency-matching could be accomplished. Within each of several fielded waves, Westat selected an equal probability sample of enumerated women **within each of the five-year age groups**, with older women having higher probabilities of selection than younger women. Each enumerated woman was selected for the sample with her pre-assigned age-specific probability. Hence, it was possible that two or more women could be selected from the same household, although this was a rare event.

See Brinton et al (1995) for a detailed description of the WISH study methodology at all three sites.

2.3 Area Probability Sampling Methodology for the Atlanta WISH Site

Selection of the area probability sample used standard techniques (e.g. Kish, 1965). No stratification was used for the area sample since the RDD sampling used no stratification. Primary or first stage sampling units were defined as block groups in the three county area (Fulton, DeKalb, Cobb) of metropolitan Atlanta, using preliminary HU estimates used by the Census Bureau in its preparation for the 1990 census (as opposed to relying on 1980 Census data). From a sampling frame of 1264 block groups, 180 sample block groups were selected using probability proportional to estimated size (ppes) sampling. Second stage sampling units were defined as segments, with a minimum of 75 estimated HUs. One segment per block group was selected with ppes sampling. The 180 segments selected for the sample ranged in estimated size from 75 to 2053 HUs.

Seven waves of 25 or 26 sample segments were defined where each county was proportionately represented in each wave. The waves were fielded sequentially in a random order. Each sample segment was mapped in the office using up-to-date detailed Atlanta maps. A field worker (counter and lister) visited each sample segment and counted the number of HUs therein. The addresses and locations of all HUs in the sample segment were listed if the sample segment contained less than about 150 HUs. Larger segments were subsegmented, with one subsegment selected with ppes sampling, and all HUs in the subsegment were listed. In the third stage of sampling a systematic random sample of about 24 sample HUs was selected from the listing sheets for each sample segment.

A female enumerator visited **all** sample HUs within a given segment to enumerate or screen all women aged 20-54 in each HU. At the time of enumeration the interviewer instituted a predetermined random selection scheme which indicated whether or not each enumerated woman was selected for the area control sample. If the selected woman (or women) was home at the time of enumeration, the interviewer attempted to conduct an interview at that time. Generally, however, the interviewer made a future appointment to conduct the interview.

The NCI/NIH funded activities discussed in Sections 2.1, 2.2 and 2.3 above resulted in the collection and computerization of the Atlanta WISH data for cases, area controls and RDD controls. All subsequent activities described in this report were conducted under the U.S. Army grant.

2.4 Weighting the RDD and Area Samples

Ordinarily, a population based control sample in a case-control study is not weighted to make inference to the population from which the sample was selected because the primary purpose of the case-control study is to compare prevalence of risk factors in controls to those in cases. However, in this methodological study it was desired to weight each of the two control samples so that both samples could be compared to each other and to the 1990 Census data. The objective of weighting each sample is to obtain a final weight for each RDD and area interviewed woman, where the value of the final weight is the number of women in the population whom the interviewed woman represents. Standard sample survey weighting procedures were used to develop the weights; these procedures are based on details of the specific sampling plan and nonresponse adjustments.

Area Sample: The selection probability for each woman in the area sample was calculated. This selection probability is the product of two probabilities: (1) the probability with which the woman's household was selected and (2) the probability with which she was selected, given that her household was selected. These probabilities are known from the area probability sampling plan. The initial weight of an area woman is the inverse of her selection probability. The initial weight was adjusted for enumeration or screening nonresponse and for interview nonresponse, somewhat inflating the value of the initial weight so that the interviewees now represent the nonrespondents as well.

The weighting process up to this point was based on the sampling frame, the sample design and internal (to the sample) nonresponse corrections. No poststratification adjustments were made to align the area sample to Census data, a common procedure in sample surveys, because it was desired to compare the area sample to the Census.

RDD Sample: It was not straightforward to weight the RDD sample from the sampling frame because the Mitofsky-Waksberg method was used to obtain an equal probability sample of residences (with a telephone). The problem is that the RDD sample itself could not be used to estimate the total number of residences with a telephone, whereas the area control sample could be (and was) used to estimate the total number of occupied housing units in the three counties. In order to develop a selection probability for each woman in the RDD sample, it was assumed that the (approximately) equal probability sample of 5464 RDD selected households (5442 households identified by Westat plus 22 additional telephone numbers allocated as households but never reached) make inference to 611,576 telephone owning households in the three counties. The point estimate 611,576 was obtained from analysis of the 1990 Public Use Microdata Sample (PUMS) Census data. Subsequent steps in the weighting procedure for RDD women

were similar to the description above for the area women. Again, no poststratification adjustments were done to the final weights for the RDD women since we wanted to compare the RDD and area samples to the 1990 Census.

The age specific selection probabilities for the RDD and area samples were not equal in all instances. Although Westat changed slightly these probabilities over time in the different waves in order to control the age distribution of RDD women selected for interview, the P.I. of this grant did not have access to the wave-specific information and could not calculate age-specific selection probabilities by wave. The inverse of the RDD age-specific selection probabilities, averaged over waves, was 113 for those 20-24 years, 47 for those 25-29, 13 for those 30-34, 4.1 for those 35-39, 2.3 for those 40-44, 1.2 for those 45-49 and 1.0 for those 50-54 (always selected for interview with certainty).

The Atlanta investigators also changed the age-specific selection probabilities over time in the different waves, not only to control the age-specific sample size but also because a programming error in the random selection procedure for women to be interviewed resulted in the selection of too many younger women in the first wave (A) of area sampling. The wave-specific selection probabilities were used in weighting the area sample, since this information was available. Hence, for the area sample, there generally is a range of values for the inverse of the age-specific selection probabilities: 150 for those aged 20-24 years, 10 to 60 (mean 26) for those 25-29, 3 to 73 (mean 11) for those 30-34, 1.4 to 25 (mean 5.2) for those 35-39, 1.0 to 5.1 (mean 1.8) for those 40-44, 1.0 to 2.4 (mean 1.4) for those 45-49, and 1.0 for those 50-54 (always selected for interview with certainly).

2.5 Definition and Comparison of Response Rates for the RDD and Area Samples

The enumeration or screening response rate is defined as the percentage of chargeable households which are successfully enumerated. A chargeable household in the area sample is a sample address which is an occupied housing unit; e.g. it is not a business or a vacant housing unit. A chargeable household in the RDD sample is a telephone number which is residential; e.g. it is not a business or a disconnected telephone number. A successful enumeration or screening is defined as determining whether or not the household contains any women aged 20-54 years and, if yes, the age of each such woman in the household with some identifying information (e.g. initials) so that future contact can be made should the women be selected for interview.

The interview response rate is defined as the percentage of eligible women who are interviewed. An eligible woman is defined as a woman selected for interview who is eligible for the study, i.e. female, resided in the three county area and aged 20-54 years. There were a few instances of selecting persons who, upon later determination, were found to be ineligible.

The overall survey response rate is defined to be the product of the enumeration response rate and the interview response rate.

A complicating factor in the RDD sample was "partial enumeration". Partial enumeration means that, during the enumeration via telephone, information was obtained that the household included women aged 20 to 54 years and the number and ages of such women in the household, but name and/or address information was not obtained for women in the household. Hence, it was difficult or impossible to contact a woman from the household if she were selected later for interview. If all the necessary contact information was obtained, the enumeration was considered complete. Many partial enumerations resulted in no women being selected for interview. For those women selected for interview from a partial enumeration, attempts were made to identify and contact the selected women. Some of these selected women could not be contacted to request an interview.

One way to calculate the enumeration response rate for the RDD sample is to consider both partial and complete enumerations as successful. A woman selected for interview from a partial enumeration who could not be located then is counted as an eligible woman who was not interviewed. On the other hand, the partial enumeration could be considered as unsuccessful, since all of the desired information was not obtained. The enumeration and interview response rates are calculated both ways for the RDD sample, i.e. counting partial enumerations as successful and then as not successful.

Chi square tests were used to compare the enumeration, interview and overall survey response rates for the area and RDD samples, assuming all sample housing units and telephone numbers to be statistically independent.

2.6 The 1990 U.S. Census Data

The 1990 Census was administered in two forms: the 100 percent (short form) and the sample (long form) Census, which requests additional and more detailed information on housing units and individuals than the short form. Use of the PUMS (Public Use Microdata Samples) 5% sample data on CD-ROM allowed comparison of women aged 20-54 in the sample Census to the two control samples (area and RDD) on the following characteristics:

- age
- race
- marital status
- place of birth (U.S. or foreign)
- high school graduation
- highest grade/degree completed
- number of live births
- household income
- presence of a telephone in the residence.

The five percent PUMS sample identifies all States and various subdivisions within them, including most counties with 100,000 or more inhabitants. Each microdata file is a stratified sample of the population, actually a subsample of the full Census sample (approximately 15.9% of all housing units) that received Census long form questionnaires. PUMS sampling was done on a housing-unit basis, with all persons within a housing unit included, in order to allow study of family relationships and housing unit characteristics. Sampling of persons in institutions and other group quarters was done on a person-basis. Vacant units were also sampled. An iterative, multi-stage procedure was used to calculate both person and housing unit weights for the PUMS ; a by-product of this weighting procedure is that housing unit or persons estimates from PUMS will for the most part be consistent with the 100 percent figures for the population and housing unit groups used in the weighting procedure (Bureau of the Census, 1992).

Most instances of missing data in PUMS were imputed, and our analyses used the imputed or allocated values provided in PUMS. The PUMS dataset contains the variable GQINST which identifies individuals residing in group quarters (both institutional and non-institutional), making it possible to exclude those individuals from analyses. We excluded group quarter residents from PUMS analyses since group quarter residents had been excluded by the sampling plans for the RDD and area controls.

2.7 Comparison of Area, RDD and Census Samples to Each Other

While many variables were measured on the **interviewed** area and RDD sample women, very limited information was available on the **selected** and/or **enumerated** (or screened) women, particularly in the RDD sample. Hence, all analyses of area and RDD samples reported here were performed only for the **interviewed** area and RDD women.

Comparisons of the three samples to each other were made using **weighted and clustered** chi-square analyses to test the null hypothesis that all three samples make inference to the same population, i.e., to women aged 30-54 residing in Cobb, DeKalb and Fulton Counties, on a given variable (e.g. age, race, etc.). The sample survey software package SUDAAN (Shah et al, 1996) was used in order to incorporate into the analyses both the weighting and clustering within primary sampling units (PSU's). Each sample was described to SUDAAN as multi-stage sampling with replacement (i.e. the finite population correction factor was ignored). The primary sampling unit (PSU) for the RDD, area and PUMS samples is the telephone bank, segment and housing unit, respectively. The three samples were concatenated into one dataset for SUDAAN analyses, where sample type (RDD, area, census) was the stratification variable. For those familiar with SUDAAN, the DESIGN was WR (with replacement) and the NEST statement included the stratification variable "sample type" and the PSU variable, as described above.

Sample sizes were adequate to allow comparisons of age by race and by county. Race was dichotomized as black and non-black due to the small number of women who reported their race as neither black nor white. Sample sizes were adequate to compare the

three samples on race, by county. Note that the selection of too many younger women in wave A of the area sample has no impact on the weighted analyses in these comparisons.

Other variables analyzed were: high school graduation (yes/no), highest educational attainment, marital status, number of live births, household income level, place of birth (US or foreign), and presence of a telephone in the residence (yes/no). For residential telephone coverage, only the area and Census samples could be compared. Sample sizes were sufficient to make comparisons for all characteristics on a county basis and to make comparisons on a racial basis for all characteristics except place of birth.

When weighted analyses began it was noted that there were very few women aged 20-29 in either the RDD, area for case groups. Further, when weighed analyses were done, the large weights for these few women tended to inflate variances. For these reasons only women aged 30-54 were used in all analyses which compared the area and RDD samples to the Census sample.

2.8 Comparison of Area and RDD Samples, Unweighted, on Census Variables

Epidemiologists typically are more interested in whether the two sample control groups are similar to each other rather than in whether the two samples make inference to the same underlying population. Hence, the area and RDD samples were compared on the same demographic and personal characteristics listed above (except for telephone coverage) in unweighted analyses. The null hypothesis tested by the chi-square test is that both the area and RDD samples make inference to the "same" population, although a somewhat "artificial" population.

In unweighted analyses, the samples do not reflect the underlying population since the selection probabilities of sampled women differ dramatically by age, with older women substantially oversampled in order to try to match the age distribution of breast cancer cases. Behaviors related to age also do not reflect the underlying population in unweighted analyses.

The unweighted chi-square tests were performed both with the clustering recognized and the clustering ignored, where the clustering variable is the PSU. Analyses which recognize the presence of clustering typically have higher variability, resulting in smaller p-values for testing null hypotheses. All of these analyses were done in SUDAAN, with appropriate instruction to SUDAAN to ignore the weighting and to recognize or not recognize the clustering.

2.9 Comparison of Area and RDD Samples on Risk Factors for Breast Cancer

Although epidemiologists are interested in whether different sampling techniques for control groups result in samples which are similar on demographic characteristics, a more salient interest is whether the sampling techniques for control groups result in comparable prevalences of the known and suspected risk factors under study. Even if the

control samples may differ somewhat on demographic factors, such as age or race, these factors typically are controlled on in epidemiological analyses which assess relationships between risk factors and disease.

In these sets of analyses the area and RDD samples were compared to each other on the prevalence of several established or suspected risk factors for breast cancer. In addition, the samples were compared on variables related to general health status. The 41 variables investigated are:

- number of live births
- age at first live birth
- number of pregnancies
- age at first pregnancy
- ever breast fed at least two months
- number of months breast fed
- number of children breast fed
- miscarriage history
- abortion history
- age at menarche
- previous breast biopsy
- body mass index
- family history of breast cancer (mother or sister)
- cigarette smoking
- alcohol consumption
- education
- ever use oral contraceptives
- years of oral contraceptive use
- years since first oral contraceptive use
- years since last oral contraceptive use
- age at first use of oral contraceptives
- income
- menopausal status
- religion
- marital status
- number of times married
- ever used IUD
- age at menopause
- ever use estrogen pills
- age at first use of estrogen pill
- on estrogen pill now
- ever prescribed medicine for high blood pressure
- number supported on income
- ever used progesterone pills

- ever had breast aspiration
- ever performed breast self exam
- ever had mammogram
- ever had high blood pressure
- ever had high cholesterol
- ever used electric blankets and/or electric mattress pads
- usual occupation

The area and RDD samples were compared in a series of logistic regression models used to predict the probability of a woman being in the RDD sample (as opposed to the area sample). The first logistic regression model included age only, dichotomized as <50 years or 50-54 years old. The second logistic model included both age and race (black or nonblack) and their potential interaction. Subsequent logistic regression models included age, race, one of the variables from the list above, and all potential two-factor interactions between these three main effects. These analyses assess whether the RDD and area samples differ significantly on any of the variables in the list above, after controlling for age and race. The logistic regression analyses were done in SUDAAN, with the two samples (RDD and area) described as simple random sampling, i.e. unweighted and unclustered. These SUDAAN analyses are identical to logistic regression analyses conducted in SAS or other statistical packages which assume simple random sampling. The Hosmer-Lemeshow goodness of fit test was assessed for the final logistic model for each model, using SAS.

2.10 Comparison of Area and RDD Samples on Odds Ratios for Breast Cancer

Another way to compare the two control samples is to determine whether the two samples yield equivalent odds ratios for risk factors of interest, generally in a logistic regression model which controls on other relevant variables. A straightforward method is to estimate a particular odds ratio using the cases and the area control sample and then estimate the same odds ratio using the cases and the RDD control sample. A comparison of these two point estimates, taking into account sampling variability, is not straightforward since the two point estimates, both based on the same cases, are statistically dependent.

An alternative analytical approach, which accounts for the statistical dependence in comparing estimated odds ratio, is to use polytomous logistic regression. The general application of polytomous logistic regression is to partition the cases into various groups and then develop logistic regression models where each case group is compared to the controls. One can test whether the effect of one risk factor variable is the same for all case groups, taking into account that the two or more estimated regression coefficients are both based on the same control group and, hence, have covariance.

In the current application of polytomous logistic regression, the dependent variable is sample type at three levels: case, area controls, and RDD controls. Here, the controls are partitioned, with only one case group being used. The polytomous logistic

regression models the probability of being an area control, compared to being a case, and the probability of being an RDD control, compared to being a case. For a given risk factor variable, it is of interest to test whether the regression coefficient is the same for the area controls as for the RDD controls. Exponentiation of the regression coefficients yields estimated odds ratios, where the odds ratios are the comparison of controls to cases. The typical odds ratio in case-control studies, obviously, is defined as cases compared to controls. Hence, the value of the estimated odds ratios in these analyses should be, roughly, the inverse of what one would expect in the typical logistic regression analyses of risk factors for breast cancer.

Each polytomous logistic regression model contains age and race as control variables. The following established or suspected risk factors for breast cancer are modeled, one by one:

- family history of breast cancer
- age at first live birth
- number of live births
- number of months breast fed
- age at menarche
- years of oral contraceptive use
- alcohol consumption
- abortion history
- miscarriage history

The null hypothesis to be tested is that the regression coefficient (or odds ratio) is the same for each of the two control groups.

3. BODY-RESULTS

3.1 Response Rates for the Area and RDD Samples

Tables 1 and 2 show the enumeration and interview outcomes for the area and RDD samples. Although only women aged 30-54 are included in most subsequent analyses, the survey response rate calculations included women aged 20-54, i.e., all women in each sample.

3.1.1 Area Sample. Table 1 shows that a total of 3804 sample HUs were selected for the area sample, of which 486 (mostly vacant units) were not chargeable (i.e. did not count against the enumeration or screening response rate). Of the 3318 chargeable HUs, 3150 were successfully enumerated (screened); thus, the enumeration (screening) response rate for the area sample was 94.9 percent (3150/3318). The most common reasons for nonenumeration were contact problems, not being at home or otherwise unavailable. Also in the area sample, 34.8 percent of enumerated HUs had no

women aged 20-54. Table 2 shows that 802 women were selected for interview, of whom 794 were eligible for interview. The interview response rate for the area sample was 80.6 percent (640/794). The most common reason for noninterview was refusal. The overall survey response rate for the area sample was 76.5 percent (.949 x .806 x 100).

3.1.2 RDD Sample. Table 1 shows that 12,033 sample telephone numbers were selected for the RDD sample, of which 6542 were not chargeable (not residential or not in the three county area). Of the remaining telephone numbers, 4927 were enumerated (4572 completely and 355 partially), 515 were not enumerated and 49 were indeterminate as to residential status.

If the 355 **partial** enumerations were counted as **successful**, the enumeration response rate was 90.2%, i.e. 4927/(5442 + (0.454 x 49)) = 4927 / 5464, where 5442 is the total number of telephones determined to be residential (4927 + 515). The term 0.454 in the denominator is the proportion of finalized telephone numbers which are residential {[5442 / (5442 + 6542)] = .0454 }. This proportion is multiplied by the number of nonallocated telephone numbers (49), to yield the number of unallocated telephone numbers that can be presumed to be working residential numbers; these are chargeable residential telephones, i.e they count against the enumeration response rate.

If the 355 **partial** enumerations were counted as **unsuccessful**, the enumeration response rate was 83.7%, i.e. 4572 / 5464, which is also equivalent to the enumeration rate above (90.2%) times the proportion of enumerations that were complete. The most common reason for RDD nonenumeration was refusal (304 of 515). In RDD sampling, 41.3 percent of the 4927 enumerated households had no women aged 20-54.

Table 2 shows that the 355 partial enumerations resulted in 42 women who were selected for interview but on whom additional contact information could not be obtained (e.g., name and/or address). Hence, these 42 women were not interviewed. Table 2 shows that some women were excluded from interview after they were selected, primarily because of incorrect enumeration data about their age, county of residence, etc. This occurred more frequently in RDD than in area sampling.

If the 355 partial enumerations were counted as successful, the 42 women selected but who could not be contacted must be included in the denominator of the interview response rate. Thus, the interview response rate would be 75.8%, i.e. 652 / (818 + 42). This would give an overall survey response rate for the RDD survey of 68.4% (.902 x .758 x 100).

If the 355 partial enumerations were counted as unsuccessful, the interview response rate would be 79.7%, i.e. 652 / 818. This would give an overall survey response rate for the RDD survey of 66.7% (.837 x .797).

Note that the second method of calculating the overall survey response rate (yielding 66.7%) will never exceed the calculation from the first method (yielding

68.4%), since the second approach counts as nonenumeration all of the 355 residences with partial enumeration, whereas the first approach counts as nonenumeration in effect only those residences from the 355 where a woman was selected for interview **and** further contact information could not be obtained.

3.1.3 Comparison of Area and RDD Samples. The enumeration response rates for the two samples were statistically significantly different no matter how the partial enumerations were handled ($p = 0.001$ for both RDD methods). The area enumeration response rate was 5 to 10 percent higher than the RDD , i.e., 95 percent compared to 84 percent or 90 percent. In addition, both the enumeration refusal rate ($p = 0.001$) and the proportion of HUs or households that could not be enumerated due to language problems ($p = 0.020$) were statistically significantly higher for RDD sampling. The RDD sample had a 6 percent enumeration refusal rate compared to 2 percent for the area sample. The proportion of HUs or households that could not be enumerated due to contact problems did not differ for the two samples ($p = 0.660$), about 3 percent for each.

The interview response rates were not statistically significantly different for the two samples ($p = 0.65$) if the RDD partial enumerations were counted as unsuccessful (79.7% for RDD vs. 80.6% for area), but were statistically significantly different ($p = 0.019$) if partial enumerations were considered successful (75.8% for RDD compared to 80.6% for area).

The overall survey response rate for the area sample was statistically significantly different from that of the RDD sample regardless of how the RDD partial enumerations were handled ($p = 0.001$ for either RDD response rate method). The overall survey response rate for the area sample (76.5%) was higher than the RDD (68.4% or 66.7%). To compare the overall survey response rates in a chi-square test, a "back-calculation" was done to obtain the number of women who "should have been" selected for interview (had there been no enumeration refusals, etc.); the number of interviewed women was divided by the overall survey response rate to provide the "denominator" of chargeable interviews for the chi-square test.

It is clear that the enumeration or screening rate for the household is significantly lower for RDD than for area sampling. By considering all partial enumerations as chargeable, the interview response rates for the two samples are comparable. The difference in the enumeration response rates between the two samples is the primary (or sole) contributor to the significantly lower overall survey response rate in the RDD sample.

3.2 Percent Ineligible Households for the Area, RDD and Census Samples

Although not specified as one of the initial objectives of the study, the area and RDD samples were found to differ in the enumeration process on identifying households as eligible for the study, i.e., households with at least one woman aged 20-54. For the purpose of investigating this further, we focus on the percentage of enumerated

households which are **ineligible** for the WISH study, i.e. contain **no** women aged 20-54 years. Table 1 shows that the percentage of ineligible households was 34.8% (100 x 1097 / 3150) for the area sample and 41.3% (100 x 2033 / 4927) for the RDD sample. A 2 x 2 chi square test of this association was statistically significant at $p = 0.001$, showing that the 41.3% ineligibility rate in the RDD sample was significantly larger than the 34.8% ineligibility rate in the area sample.

Because this finding was interesting and , to our knowledge, has not been demonstrated before, we used more conservative statistical techniques to assess whether the finding held up. The chi-square test performed in the above paragraph, based on the figures in Table 1, does not take into account the clustering or weighting of households in either the area sample or the RDD sample, although Tables 1 and 2 use standard procedures for reporting response rates in field work (i.e. unweighted).

Thus, we used the area sample, the RDD sample, as well as the Census PUMS data, to estimate the percentage of housing units in the three county area which contain no women aged 20-54 years. The Census PUMS data was analyzed using the appropriate weight for each housing unit and recognizing any clustering in the dataset. The area sample was analyzed similarly, using the weight for each sample housing unit with the sample housing units clustered into segments.

Unfortunately, we did not have access to Westat data which would allow us to classify the RDD identified housing units of Table 1 into their appropriate telephone bank (i.e. cluster). The design effects due to Mitofsky-Waksberg clustering tend to be smaller than the design effects associated with area sampling, as borne out in our WISH RDD and area datasets as well. Further, the "weights" associated with the RDD identified housing units are all approximately equal, even though they are not explicitly calculated as part of the Mitofsky-Waksberg sampling procedure. Hence, assuming the RDD sample to be a simple random sample of housing units, for the purpose of estimating the percentage of ineligible housing units in the three county area, is not expected to underestimate the standard error of the point estimate by a substantial degree. For comparison purposes, however, a design effect of 2.0 was assumed for the RDD sample to see whether the observed finding still held up under some assumed degree of clustering of telephone numbers.

Table A below shows that the RDD sample estimates a substantially larger percentage of ineligible housing units (41.3%) than either the area sample (34.9%) or the 1990 Census PUMS data (33.5%). Even with assuming a design effect of 2.0 for the RDD analysis, a conservative position making it harder to show a statistically significant difference between RDD and the other two samples, the RDD 95% confidence interval on percentage of ineligible housing units falls totally outside the 95% confidence intervals based on the area and census samples. Note that the 95% confidence interval based on the area sample includes the 95% confidence interval based on the census sample, indicating that the point estimate from the area sample is consistent with the point estimate from the census sample. Finally, a linear contrast comparing the average of the

area and census point estimates to the RDD point estimate was statistically significant (z = 5.33, p < .0001), indicating that the RDD sample, compared to the other two samples, estimates a significantly larger percentage of ineligible housing units in the metropolitan Atlanta three county area.

TABLE A

Percentage of Housing Units Having No Women Aged 20-54 Years by Four Estimation Techniques, Metropolitan Atlanta, 1990-92

Estimation Method	Point Estimate	Standard Error	95% Confidence Interval
1990 Census PUMS	33.5%	0.29%	(32.9%, 34.0%)
Area Sample	34.9%	1.74%	(31.5%, 38.3%)
RDD (as simple random sample)	41.3%	0.70%	(39.9%, 42.6%)
RDD (assumed design effect of 2.0)	41.3%	0.99%	(39.3%, 43.2%)

3.3 Weighted and Clustered Comparisons of Area, RDD and Census Samples

The intent of these analyses was to assess whether the area and RDD samples made statistical inference to the same population as the 1990 Census PUMS sample for women aged 20-54 years in Cobb, DeKalb and Fulton counties in metropolitan Atlanta. The 1990 Census PUMS data can be considered as a "gold standard" since it is based on a much larger sample size and the Census Bureau expended much more resources than the WISH study to obtain complete coverage and high response rates.

3.3.1 Age. The estimated age distributions given by the three samples were not statistically significantly different over all races and counties together (p = 0.60) or when the comparisons were performed by county (p = 0.63 for Cobb, 0.88 for DeKalb and 0.12 for Fulton) or by race group (p = 0.10 for Blacks and 0.58 for non-Blacks). Hence, the **interviewed** RDD and area samples and the Census sample all make inference to the same population as far as age distribution is concerned. The age distributions are not presented in a table since they do not differ; this system is used throughout this section.

3.3.2 Race. Race was considered as a dichotomous variable: Black and non-Black. The estimated proportions of non-Black women given by the three samples are shown by county in Table 3. The estimated race distributions given by the three samples did not differ statistically over all three counties together (p = 0.17), or in DeKalb (p = 0.94) and Fulton (p = 0.12) Counties, but did differ statistically for Cobb County (p = 0.01). For Cobb County, the area sample yielded an estimated race distribution that is

quite similar to that given by the sample Census, while the RDD sample estimated about 5 percent more of the Cobb County population to be non-Black. Note the consistent pattern in Table 3 where the RDD sample estimates a higher percentage non-Black than either the area sample or the Census sample, over all counties and within each county.

3.3.3 Place of Birth. The estimated proportions of women born in the U.S. given by the three samples are shown in Table 4. Due to the sample sizes, this analysis could be done by county but not by race group. The Chi-square test found that estimates of proportion of U.S. born women given by the three samples differed statistically over all three counties together ($p = 0.02$) and in Cobb County ($p = 0.03$), marginally in DeKalb County ($p = 0.054$) and not in Fulton County ($p = 0.68$). With the exception of Fulton County, note that both the area and RDD samples estimated a higher proportion of U.S. born women than the sample Census.

3.3.4 Education. The estimated percentages of women who graduated high school given by the three samples are shown in Table 5 by county and in Table 6 by race group. In the by county analysis, the estimated percentage of high school graduates differed statistically over all three counties together ($p < 0.01$) and in all counties separately except Fulton ($p = 0.02$ for Cobb, 0.01 for DeKalb, and 0.08 for Fulton). Note that the estimates from the Census were consistently the lowest and those from the RDD sample were consistently the highest.

The estimated high school graduation rates also differed statistically among the three samples, stratified by race, i.e., for Blacks ($p < 0.01$) and non-Blacks ($p = 0.02$). Both the area and the RDD samples overestimated the percentage of high school graduates compared to the sample Census estimate, with the area sample generally giving estimates closer to the Census than the RDD sample. As with the analyses by county, the Census consistently gave the lowest estimates and the RDD sample consistently gave the highest estimates.

When educational attainment was assessed as the highest grade or degree completed (less than high school graduation, high school graduation, post-secondary training but less than a Bachelor's degree, Bachelor's degree, higher degree than Bachelor's), a statistically significant difference was found between the estimated distributions of highest grade or degree completed given by the three samples when all three counties and all races were considered together ($p < 0.01$) but not for individual counties ($p = 0.09$ for Cobb, 0.16 for DeKalb, and 0.43 for Fulton) or races ($p = 0.07$ for Blacks and 0.27 for non-Blacks). In the analysis over all races and all counties, both the area and the RDD samples estimated a somewhat higher educational level than the sample Census. For example, the proportion of the population estimated to hold at least a Bachelor's degree was 32.8 percent, sample Census; 37.1 percent, RDD sample; and 37.6 percent, area sample.

3.3.5 Marital Status. Tables 7 and 8 show the estimated distributions of marital status (currently married; widowed, divorced or single (never married); and separated)

given by the three samples by county and by race respectively. The estimated distributions differed statistically by sample type over all three counties together ($p = 0.01$) and in Cobb ($p = 0.03$) and Fulton ($p = 0.01$) Counties individually, but not in DeKalb County ($p = 0.50$). The estimated marital distributions among the three samples also differed statistically for non-Blacks ($p < 0.01$) but not for Blacks ($p = 0.51$). For non-Black women, both the area and RDD samples estimated a statistically significantly higher proportion of women who were currently married and a statistically significantly lower proportion of women who were either widowed, divorced or single or who were separated from their spouses than the sample Census.

In the analysis over all races and all three counties together, both the area and the RDD samples overestimated the proportion of women who were currently married and underestimated the proportion who were either widowed, divorced or single compared to the sample Census. This same pattern of differences from the Census sample estimates was observed in analyses of all races in Fulton County and for non-Blacks over all three counties.

3.3.6 Number of Live Births. The estimated distributions of the number of live births experienced by women (none, one, two, three, four or more) given by the three samples did not differ statistically over all races and counties ($p = 0.76$) or for individual counties ($p = 0.98$ for Cobb, 0.67 for DeKalb, and 0.79 for Fulton) or for races ($p = 0.94$ for Blacks and 0.46 for non-Blacks).

3.3.7 Household Income. The estimated household income (< \$20,000; \$20,000-\$34,999; \$35,000-\$49,999; \$50,000-\$69,999; $\geq$ \$70,000) distributions yielded by the three samples did not differ statistically for all races and counties together ($p = 0.75$) or for individual counties ($p = 0.47$ for Cobb, 0.81 for DeKalb, 0.19 for Fulton) or for races ($p = 0.65$ for Blacks and 0.87 for non-Blacks). This was the only characteristic discussed to this point for which there was missing data in WISH (refusals, don't knows); 26 (4.1%) RDD controls and 18 (2.9%) area controls did not supply information on household income.

3.3.8 Residential Telephone. The estimated percentage of women living in households with a residential telephone could be compared to the sample Census estimate for the area sample only. The estimates of residential telephone coverage from the two samples by county and by race are shown in Tables 9 and 10 respectively. When analysis was performed by race group, no statistically significant difference was found between the telephone coverage estimates from the two samples for either Blacks ($p = 0.99$) or non-Blacks ($p = 0.90$). When analysis was performed by county, no statistically significant difference was found between the area sample and the sample Census estimates for all three counties together ($p = 0.92$) or for the three counties separately (DeKalb, $p = 0.69$; Fulton, $p = 0.77$; Cobb, not testable).

Hence estimated residential telephone coverage was equivalent for the area and Census samples, about 97 percent overall. Telephone coverage seemed somewhat higher

in Cobb County and slightly lower in Fulton County, most likely due to a smaller percentage of Blacks in Cobb County and a much larger percentage of Blacks in Fulton County. Telephone coverage was estimated as 99 percent for non-Blacks and 94 percent for Blacks.

3.4 Unweighted Comparisons of Area and RDD Samples

These analyses compare the area and RDD samples to each other, from the viewpoint of an epidemiologist who typically would not do a weighted analysis. Although an epidemiological analysis typically would not recognize the clustered nature of the data in either the RDD or area sample, we present here standard errors and tests of statistical analysis for both the unclustered and clustered approaches.

3.4.1 Age. The unweighted age distribution of the two samples by race are shown in Table 11; standard errors are included for both clustered and unclustered analyses. In both the clustered and the unclustered analysis, Table 11 shows statistically significant differences between the unweighted age distributions of the two samples over all three counties together ($p = 0.01$ for clustered and unclustered). The RDD sample was older than the area sample. When area and RDD age distributions were compared by race group, statistically significant differences were not found for Blacks or non-Blacks separately because of the reduced sample size, although the age distributions, by race, continued to show the pattern of an older age distribution in the RDD sample. This pattern is not surprising because younger women were inadvertently selected with too high a probability at the beginning of the study for the area sample. Detailed comparisons of the age distribution of the area and RDD sample, by wave (not shown here), confirm that the only reason for the different unweighted age distributions in the area and RDD samples is due to the implementation of the area sampling plan, whereby too many younger women were selected in wave A (the first wave fielded).

3.4.2 Race. The unweighted race distributions of the two samples, including standard errors from both the clustered and unclustered analyses, are shown in Table 12. Statistically significant differences in the racial distribution of the two samples were found in unclustered analyses over all three counties together ($p < 0.01$) and in each county individually ($p = 0.01$ for Cobb and DeKalb and 0.04 for Fulton), with the RDD sample having about 10 percentage points more non-Blacks than the area sample. In the clustered analyses, however, no statistically significant differences were found in the race distribution of the two samples over all counties or in any individual county.

The difference in the statistical significance of the clustered and unclustered analyses is due to the difference in the size of the standard errors. The standard errors for the clustered analyses are larger because the intracluster correlation is high for the variable race; that is, race is not randomly distributed within clusters (segments or telephone banks) because of housing patterns in metropolitan Atlanta.

Note also in Table 12 that the size of the standard error is greater for the area clustered analyses than for the RDD clustered analyses, even though the sample size for the two control samples is approximately the same. This could be due to a larger value of the intracluster correlation coefficient (for race) in the area sample and/or due to a larger PSU (primary sampling unit) or cluster size in the area sample (mean of 4.67 interviewed women per segment in the area sample versus 2.03 interviewed women per telephone bank in the RDD sample). An initial investigation to determine which of these two factors was most important (PSU or cluster size vs. magnitude of intracluster correlation coefficient) was not successful and was not pursued further since the issue is tangential to the specific aims of the Army grant.

Because the RDD and area samples obviously differ on age (unweighted), and because age is related to race (proportion Black is somewhat higher among the younger population in metropolitan Atlanta), one possible explanation for the observed difference in race seen in Table 12 is the difference in age between the control samples. To test this hypothesis, an age stratified, unweighted and unclustered test for the independence of race and sample type was performed. The summary (1 df) Cochran-Mantel-Haenszel statistic indicated an overall statistically significant relationship between sample type and race ($p = 0.002$). This relationship was statistically significant only for the 40-45 year age group ($p = .028$), not for the other three (30-39, $p = 0.091$; 45-49, $p = 0.168$; 50-54, $p = 0.428$) and in the same direction in all age groups. Thus, the difference in age distributions between the area and RDD samples is not the explanation for the difference in the race distributions.

The conclusion is not clear cut as to whether the area and RDD samples, unweighted, differ on race. In unclustered analyses the two samples are significantly different on race, even after controlling for the known age difference between the two samples. The race distributions look quite different, with Table 12 indicating that the area sample has a higher percentage Black by 7 to 12 percentage points. However, when clustering is taken into account, the two control samples no longer differ significantly on race, although the p-value over all three counties is $p = .09$. Because the design effect for race is so high in these datasets, due to housing patterns in Atlanta, and because of evidence from other analyses to be described later, we conclude here that the two samples RDD and area, do differ on race in addition to age.

3.4.3 Place of Birth. The percentage of U.S. born women in the two samples did not differ statistically in either clustered or unclustered unweighted analysis. There was no statistically significant difference over all three counties or in any county individually.

3.4.4 Education. There were no statistically significant differences in the percentage of women who graduated high school or in the distribution of highest grade or degree completed in either the clustered or the unclustered analysis. No statistically significant differences were found over all counties and races, or for individual counties or race groups.

3.4.5 Marital Status. Neither the clustered nor the unclustered analysis found a statistically significant difference in the two samples on distribution of marital status. No statistically significant differences were found over all counties and races, or for any county or race group individually.

3.4.6 Number of Live Births. No statistically significant differences were found in the number of live births reported by women in the two samples in either the clustered or the unclustered analysis. Statistically significant differences were found neither in the by county or by race analysis, nor in the analysis over all counties or all races.

3.4.7 Household Income. Statistically significant differences in the distribution of household income level were not found in either the clustered or the unclustered analysis. No statistically significant differences were found over all counties or all races, or for individual counties or race groups.

3.5 Breast Cancer Risk Factors in Area and RDD Samples

The base logistic regression model for these analyses contained two independent variables: age (dichotomized as < 50 or 50-54, with age 50-54 as the reference group) and race (dichotomized as Black and non-Black, with non-Black as the reference group). The dependent variable was type of control group (area or RDD). The base logistic regression model predicted the probability of the woman being in the RDD control group. All analyses in this section are unweighted and unclustered. All area and RDD controls who reported a previous history of breast cancer were excluded from these analyses, reducing the sample size slightly from a total of 1292 controls (640 + 652) to 1272 controls.

TABLE B

Estimated Regression Coefficients, Odds Ratios and p-values
for Predicting Probability of Being in RDD Control Group

Regression Coefficient	Point Estimate	Odds Ratio	p-value
Intercept	.4173		
Age	-.3989	0.67	.0024
Race	-.3450	0.71	.0049

The two-way interaction between age and race was not statistically significant in the base logistic regression model, and the main effects model with age and race had a very good fit according to the Hosmer-Lemeshow goodness of fit test (p = .82). Table B above gives the estimated regression coefficients, and odds ratios, for the base model. The interpretation of the regression coefficients and the odds ratios is that younger women, compared to older women, are less likely to be in the RDD control group and

that Black women, compared to non-Black women, are less likely to be in the RDD control group. This base logistic regression model indicates that age and race are both strongly, and independently, associated with type of control group; this finding is consistent with the other unweighted and unclustered analyses reported in the preceding paragraphs of this Results section.

Each of the forty-one variables (breast cancer risk factors and general health status indicators) listed in Section 2.9 (pages 16-17) was added, individually, to the base logistic regression model above, along with the possibility of all two-factor interactions of age and race with the added variable. The best fit was obtained for the logistic regression model with age, race and the one added variable. Then, the statistical significance of the added variable was tested, conditional on age and race being in the model. If the added variable was not statistically significant in the model, this was interpreted as the area and RDD control groups not differing on the added variable, after controlling on age and race.

The remarkable finding is that 34 of the 41 variables investigated were statistically nonsignificant ($p > .05$), conditional on age and race already being in the logistic regression model. All of these logistic regression models had a Hosmer-Lemeshow goodness of fit p-value ranging from .18 to .98, with most of the goodness-of-fit p-values being around .8 or .9.

Of the seven variables which were statistically significant at a p-value close to .05, five were related to oral contraceptive use and the other two were history of breast biopsy and family history of breast cancer. Although these seven variables are discussed below, the main message is that some findings would be expected simply by chance, given that 41 different models were fit. Further, these few significant findings do **not** warrant a strong conclusion that area and RDD sampling yield control groups with different characteristics on breast cancer risk factors and general health status.

The five variables related to oral contraceptive (OC) use, and their associated p-value in their respective logistic regression model, are:

- ever use OC more than 6 months (yes, no), $p = .0585$
- total years used OC (3 groups, with no use as referent), $p = .0550$
- years since first OC use (3 groups, with no use as referent), $p = .0489$
- years since last OC use (3 groups, with no use as referent), $p = .0379$
- age at first OC use (2 groups, with no use as referent), $p = .0245$

The estimated odds ratio for the first logistic regression model listed above (ever use OC) was .75, indicating that ever users of oral contraceptives were somewhat less likely to be in the RDD control group. All of the estimated odd ratios in the remaining four models for other aspects of OC use were remarkably close to .70, indicating that the two control groups do not differ significantly on any of the more finely measured aspects of OC use (i.e. total years, years since first or last use, age at first use). The five logistic regression

models for the OC variables consistently indicate that ever users of OC are somewhat, marginally, less likely to be in the RDD control group. The five p-values indicated above are not very small, and three of the five and not even less than .05. Hence, the conclusion is that the area and RDD control groups do not differ significantly on aspects of OC use, after controlling for age and race, though there is a slight suggestion that that OC use may be more common in the area control group. Further analyses underway, not included in this final report, are investigating this issue more closely.

The two control groups differed marginally on history of a breast biopsy (p = .0448), controlling on age and race, with history of biopsy increasing the odds of being in the RDD control group (odds ratio = 1.42).

Finally, the two control groups differed significantly on history of breast cancer (p = .0126), controlling on age and race, with family history decreasing the odds of being in the RDD control group (odds ratio = .55).

Of the seven variables (out of 41 investigated) that seemed to perhaps discriminate between the area and RDD control groups, the only one with a p-value close to .01 was family history. Given the multiple testing done in this section (41 different models), finding one seemingly important variable would be expected by chance. Hence, as stated above, the general conclusion is that the two control groups do not differ substantially on breast cancer risk factors or on health status variables, after controlling on age and race.

3.6 Odds Ratios in Area and RDD Samples

These analyses are in progress, and the submission of the final report could not be delayed further to include these results. Given the many results discussed above, it is anticipated that the estimated odds ratios for breast cancer risk factors will be substantially the same for the two different control samples, after controlling on relevant demographic and other factors in the analysis. These odds ratio results will be presented at the Army "Era of Hope" meeting in October of 1997 and will be included in one of two manuscripts being prepared from the results reported above.

4. BODY-DISCUSSION

4.1 Survey Response Rates

The enumeration or screening response rate was lower for the RDD sampling methodology than for the area sampling methodology—by about 10 percentage points if the incomplete enumerations are counted as chargeable. Under this scenario, the interview response rates were equivalent for the two samples. Hence, the overall survey response rate was significantly lower for the RDD sample, compared to the area sample.

One major reason for the lower enumeration or screening response rate in the RDD sample is the higher refusal rate (to enumeration) in the RDD sample, 5.6% for RDD compared to 1.7% for the area sample. It is easier to refuse to participate in the enumeration phase over the phone than refuse in person to an interviewer and/or enumerator who is standing at the front door. A second reason for a higher enumeration refusal rate in RDD sampling is that the household informant, particularly if female, may be more reluctant to reveal the composition of the household over the telephone than face-to-face, where the inquirer can be asked to show credentials about the research project and/or the interviewer.

The other major reason for the lower enumeration response rate in the RDD sample was the occurrence of incomplete enumerations, i.e. failure to obtain identifying information (name, initials, address) on the women aged 20-54 within the household, once it was determined that there was, indeed, at least one woman aged 20-54 in the household. Incomplete enumerations did not occur in the area samples. Again, this probably is due to reluctance on the part of the enumeration household respondent to reveal identifying information over the telephone about the particular females aged 20-54 in the household.

Unfortunately, it was not possible to ascertain whether the real difference in the enumeration or screening response rates between the two samples was related to some characteristic of the household, e.g. race, household composition, or socio-economic status. We have virtually no information on households which did not respond to the enumeration or screening, and, in addition, very limited information on those households which did respond at the enumeration and screening phase. In both the RDD and area samples, there was a deliberate attempt to ask the minimum number of questions possible at enumeration and screening in order to increase the response rate at this stage.

In this era of increasing concern about one's personal security in society, it is not surprising that household respondents may be unlikely to reveal detailed knowledge about their household composition, particularly when they are questioned about younger (20-54) female occupants. The data in this study support this supposition.

4.2 Percentage of Housing Units Which Are Ineligible

RDD sampling estimated a significantly higher percentage of housing units in metropolitan Atlanta to have no women aged 20-54 years (41.3%), compared to either the area sample or the 1990 Census PUMS data. The area sample and the PUMS data were consistent in their estimation of percentage of ineligible housing units, about 34%.

One reason that the RDD percentage is higher may be the greater reluctance of household respondents, perhaps especially women, to reveal over the telephone that their household actually contains one or more women aged 20-54 years. A household respondent may feel safer revealing this information to an enumerator or interviewer at the front door who can present credentials to review. In addition, it may be harder to hide

the presence of a female household occupant aged 20-54 in a face-to-face enumeration than over the telephone, especially if the household respondent to the enumeration is a female in this age range. Hence, it may not be surprising that the percentage of eligible households found by area enumeration is higher, compared to RDD enumeration.

Note that failure of the enumeration respondent to answer honestly about the composition of the household is not the same thing as refusing to provide any enumeration information at all, which was discussed above. A household respondent may not admit over the telephone that there are actually women aged 20-54 in the household. Thus the enumeration is successful, but the household is classified as ineligible rather than as eligible, which raises the percentage of ineligible households. Note also that this action (denying that eligible women reside in the household when in fact they do) either has **no impact on the enumeration response rate or increases the enumeration response rate** because it is a successful enumeration.

It seems quite clear from the comparison of the RDD sample to the area and Census samples that the RDD sampling methodology in this instance missed the identification of a significant percentage of eligible households in the enumeration or screening process. This increases the cost of the RDD screening procedure, although it still most likely is less expensive than using area sampling for screening and enumeration. The more important question, though, is whether the households which are missed, i.e. do not identify themselves as eligible when they actually are, differ in some systematic way from the eligible households who actually do identify themselves as such.

Unfortunately, this study has no direct information to determine whether there is some selection bias evident in the households who identify themselves as eligible during the RDD enumeration or screening phase. However, based on the extensive comparisons between the RDD and area samples conducted in this research, there seems to be no compelling evidence that such a selection bias operated in the WISH study.

4.3 Comparison of Area, RDD and Census Samples—Weighted, Clustered Analyses

In weighted and clustered analyses, both the area and RDD samples of women make similar and reasonable inference to the sample Census on age, number of live births, household income, and telephone coverage (area sample only).

A comparison of the three samples, weighted and clustered, on race produced equivocal results. Table 3 shows a pattern where the RDD sample always estimates the largest percentage non-Black, county specific and over all three counties. The area sample race distribution is closer to the sample Census race distribution than is the RDD race distribution. However, most of the comparisons of RDD, area and Census samples on race distribution do not reach statistical significance, most likely because of the large design effect for race in the area and RDD samples. Also, some people believe that the 1990 Census, and hence the sample Census, underestimates the percentage Black, particularly in urban areas. Hence, although there is no compelling evidence that the

three samples (area, RDD and Census) differ on their race distribution, there is a suggestion that the RDD sample estimates a larger percentage of the population to be non-Black.

Both the area and the RDD samples estimated a higher percentage of women born in the U.S. than did the Census. However, these differences may not be of practical significance since, over all three counties, the Census estimates 94 percent U.S. born while the area and RDD samples estimated 95 and 96 percent respectively.

Education is another characteristic on which the three samples differed, as well as for Blacks and for non-Blacks separately. In every comparison, the RDD sample estimated the highest high school graduation rate and the Census estimated the lowest rate. Two possible reasons for these differences are interview response bias and RDD sampling bias. Since both the area and RDD samples yielded consistently higher estimates than did the Census, it would appear that women with less than a high school education were less likely to agree to be enumerated or interviewed than women with at least a high school education. The fact that the RDD sample consistently yielded the highest estimates further suggests that women residing in households with telephones are more likely to be high school graduates than women living in nontelephone households. This supposition is consistent with findings from studies of telephone coverage. As seen in Table 6, the differences between the area and RDD samples in estimated high school graduation rate are greater for Blacks. Over all three counties together, the RDD sample estimate for high school graduation rate is 8 percent higher than the Census estimate for Blacks but only 2 percent higher than the Census estimate for non-Blacks. One possible explanation for the larger effect among Blacks is that the high school graduation rate is not as high among Blacks, so there is less of a ceiling effect.

The three samples also differed on marital status. Statistically significant differences were found over all counties and races together and for non-Blacks. In county based analyses, the area and RDD samples estimated a higher percentage of currently married women and a lower percentage of women who were either widowed, single or divorced than did the Census. This pattern was also seen for non-Black women, but not for Black women. Perhaps a greater reluctance of single, widowed or divorced women to be enumerated and/or interviewed could explain at least in part the differences in estimates of marital status by type of samples.

Looking at the point estimates for percentage non-Black and percentage born in the US over all counties together and for percentage of high school graduates and percentage currently married over all counties and races together, the point estimates from highest to lowest exhibit the pattern RDD/area/Census, except for place of birth. This pattern is consistent with what might be expected for characteristics that are related to telephone coverage and/or willingness to be enumerated/interviewed. For example, since telephone coverage increases with level of education, the RDD sample, compared to the Census, would be expected to predict a higher proportion of women who are at least high school graduates. Also, if willingness to be enumerated/interviewed increases with level

of education, then the area sample, compared to the Census, would be expected to predict a higher proportion of women with at least a high school education, but not as high a proportion as the RDD sample predicts. Results on applicable characteristics for which no statistically significant differences between samples were found in the weighted and clustered analysis (highest educational attainment (at 5 levels) and income) were also examined for this pattern of point estimates. The RDD/area/Census pattern was found only for highest educational attainment, where the pattern of point estimates for those having at least a Bachelor's degree is consistent with that for high school graduates.

Thus, in weighted and clustered analyses, the area and RDD samples were similar on many characteristics, but did differ from the Census on the proportion of U.S. born women, high school graduation rate, and marital status distribution. These differences most likely are due to response patterns. In addition, there is a suggestion that the RDD sample may estimate a lower percentage of the population which is Black; this may be due to phone coverage and/or due to response patterns also.

4.4 Comparison of Area and RDD Samples--Unweighted

Comparisons of the two samples in unweighted analyses were performed because the epidemiologist or case-control analyst is interested in whether area and RDD sampling procedures, both designed to select controls that are age frequency matched to cases, yield comparable samples. The two samples did not differ significantly on most of the Census variables investigated. These variables are place of birth, high school graduation, highest grade completed, marital status, number of live births and household income.

It is clear that the area and RDD samples **do** differ on age in **unweighted** analyses and **do not** differ on age, with each other or with the 1990 Census, in **weighted** analyses. The area sample had a higher percentage of women who were 30 to 39 years of age and a lower percentage of women who were 50-54 years of age (unweighted). An investigation of age-specific interview nonresponse rates for the area and RDD samples revealed no major differences between the two samples. Further, if there had been significant differential interview nonresponse by age, the comparison of the area, RDD and Census samples via weighted analyses would have shown either the RDD or area sample, or both, to differ significantly from the Census sample. Thus, it seems clear that the only reasonable explanation for finding that the area and RDD samples differ on age in unweighted analyses is the inadvertent inclusion of a larger proportion of younger women in the area sample, compared to the RDD sample. Hence, this finding of age differences between the area and RDD samples does **not** imply possible differences between area and RDD sampling procedures.

It is also clear that the area and RDD samples differ on race in unweighted and unclustered analyses, with a higher percentage of black women in the area sample. This race difference in unweighted analyses persists after controlling for age. The two samples differ marginally on race in unweighted and clustered analyses and in weighted and

clustered analyses. Especially in the area sample, the design effect for estimating percentage of women black is quite high, making it difficult to achieve a statistically significant race difference between the two samples with a very small p-value. Relatively high design effects are not surprising in this situation, given the housing patterns in metropolitan Atlanta. Taking all evidence into consideration, it seems that the area and RDD samples probably **do** differ on race, with a somewhat larger percentage of black women in the area sample.

However, what is not so clear is whether this race difference between the RDD and area samples, after controlling for age, is due to the difference in sampling techniques. A possible limitation of this study is that all analyses based on the RDD and area samples are based on **interviewed** women. Clearly, these two samples differ on method of sampling, providing the basis for this unique research project. However, it is possible that the two samples may differ, in addition, due to differential enumeration or screening nonresponse as well as differential interview nonresponse, especially with respect to race. It is also possible that the RDD identification of fewer eligible households occurs with a higher probability among black households than among white households, resulting in a higher percentage of Blacks in the area sample.

In the area sample, the interviewers were instructed to record the most likely race (based on interviewer observation) of anyone who was contacted at the household, even though the household did not complete the enumeration and screening. The area interviewers also were instructed to record the most likely race (based on interviewer observation) of the person who completed the enumeration and screening for the area sample. With such information recorded, it would have been possible to assess whether the enumeration or screening response rates differed, by race, in the area sample. Clearly, such race information was not possible to assess by interviewer observation in the RDD sample. Unfortunately, this race information was not recorded systematically by the area sample interviewers because it was not asked explicitly on the enumeration form.

Furthermore, there was no systematic data collection on the race of a woman selected for interview, in either the area or RDD samples. Hence, it was not possible to determine race-specific interview nonresponse rates. Recall that the interview response rate was quite similar for the area and RDD samples.

It is possible that area sampling results in a larger percentage of Black women because of the presumed lower telephone coverage of Blacks, compared to whites, in the southeast. Table 10 estimates that 94% of Black women aged 20-54 years and 99% of white women aged 20-54 in this three county area live in a household with a residential telephone. Hence, a presumed lower telephone coverage rate among black households does not seem to be the only explanation for the race difference in the two samples.

Thus, in unweighted analyses that incorporate clustering, the area and RDD samples are the same on all characteristics investigated except age distribution. The significantly younger age distribution in the area sample is **not** due to differences in

sampling methods. As noted in the weighted and clustered analyses, there is a definite suggestion that the area sample has a higher percentage of Black women, which is statistically significant in unclustered analyses but not in clustered analyses.

4.5 Comparison of Area and RDD Samples on Risk Factors for Breast Cancer

In the logistic regression modeling, with type of control sample as the dependent variable and analyses which were unweighted and unclustered, it was clear that the two control samples differed substantially on age and race. The actual and potential reasons for this were discussed above.

A strong finding from these analyses was that the two control samples, area and RDD, of interviewed women did not differ substantially on breast cancer risk factors nor on several other variables related to general health status, after controlling on age and race in the logistic regression model. This finding is of great importance to the breast cancer epidemiological community since the empirical data presented here indicate that RDD sampling for control group acquisition does not seem to have any serious biases, compared to area sampling. The analyses performed in this research project looked thoroughly for potential substantive differences between the two control groups, but did not find any major differences.

4.6 Cost of Area and RDD Sampling

A major difference between area and RDD sampling is the cost of selecting the samples; area sampling tends to be more expensive than RDD sampling. In the WISH study, the overall cost differential would not be expected to be as great since all interviews in both control groups were conducted in person; thus the interviewing costs were the same for either sampling method. Although it would be of interest to quantify the RDD and area sampling costs for the WISH study, it is not possible to do so since exact cost data were not recorded for the sampling activities.

4.7 Limitations of the WISH Dataset

The WISH dataset, although unique in its capability to compare area and RDD sampling, does have some limitations. The data were collected from one geographic site (three counties in metropolitan Atlanta) and, essentially, only on women aged 30-54 years old since there were so few women aged 20-29 in the samples. Hence, the conclusions on comparing RDD with area sampling may not apply to rural areas and other demographic groups. However, since it is so difficult and expensive to conduct a case-control study with two control groups such as the WISH study has done, this unique dataset yields valuable information that addresses some of the lingering concerns regarding RDD sampling in breast cancer case-control studies.

5. CONCLUSIONS

1. RDD sampling has a significantly lower enumeration or screening response rate than does area sampling, and, hence, also has a significantly lower overall survey response rate. The lower enumeration response rate in RDD sampling most likely is due to the fact that it is easier for a household respondent to refuse over the telephone than to refuse someone who is at the front door of the household.

2. RDD sampling finds a significantly higher percentage of households with no women aged 20-54 years in the enumeration process, compared to area sampling and to the 1990 Census PUMS sample. This seems almost certainly to be due to the fact that household respondents to the RDD enumeration/screening process may not admit that there are women aged 20-54 in the household. Although no data were available to directly investigate whether the RDD households who admitted to having women aged 20-54 years had some sort of selection bias, the many analyses undertaken in this research project do not point to any potentially severe selection biases due to the RDD underidentification of households with younger women.

3. In weighted and clustered analyses, the area and RDD samples make similar and reasonable inference to the same population as does the sample Census on many of the characteristics included in the Census data. Differences that do exist for some characteristics, such as education and marital status, probably are due more to enumeration and interview response patterns than to a difference in area and RDD sampling methods. These analyses provide evidence of the comparability of the area and RDD sampling methodologies with respect to the limited variables available in the Census data.

4. In unweighted analyses, the area and RDD samples differed significantly on age, but this was due to an error in wave A of the area sample when too many younger women were inadvertently selected for interview. Hence, this age difference in the two control samples, with the area sample being younger, is not a function of differences between area and RDD sampling methodology.

5. In unweighted analyses, the area and RDD samples differed significantly on race, even after controlling on age. The area sample had a higher percentage of Blacks, compared to the RDD sample. It is not definitely clear from the data available to us why this occurred. Possible reasons are somewhat lower telephone coverage among Blacks and possible race related response patterns to the enumeration or screening procedure and/or to the interview itself. Hence, the difference in the race distribution between the area and RDD samples possibly is related to the two different sampling methodologies.

6. In unweighted analyses of one variable at a time, no differences were found between the area and RDD samples on several characteristics that are "known" risk factors for breast cancer: education, marital status, number of live births and household income.

This is evidence that the two sampling methodologies have resulted in somewhat equivalent control groups.

7. In logistic regression modeling (unweighted, unclustered) to assess whether the area and RDD samples differ significantly on breast cancer risk factors or on general health status variables, after controlling for age and race, there is a preponderance of evidence which indicates that the two control samples do not differ substantially on the 41 variables investigated. This is also evidence that the two sampling methodologies have resulted in fairly equivalent control groups.

8. In this study, conducted at one geographic site (metropolitan Atlanta) and on a specific subpopulation (women aged 20-54 years), area and RDD sampling yield very comparable control group samples for a breast cancer case-control study. The two control samples are comparable both in terms of weighted analyses which make inference to the larger population and in terms of comparing the two samples in unweighted (clustered or unclustered) analyses. The few characteristics on which differences are present between the two unweighted samples, i.e. age (due to known reasons) and race (due to unknown reasons, but speculated), would certainly be controlled on in standard epidemiological analytical methods to investigate potential risk factors for breast cancer.

6. REFERENCES

1. Aquilino WS, Losciuto LA. Effects of interview mode on self-reported drug use. *Public Opinion Quarterly* 1990; **54**:362-95.

2. Biemer P, Akin D. The efficiency of list-assisted random digit dialing sampling schemes for single and dual frame surveys. *Proceedings of the American Statistical Association, Section on Survey Research Methods* 1991: 1-10.

3. Blair J, Czaja R. Locating a special population using random digit dialing. *Public Opinion Quarterly* 1982; **46**:585-590.

4. Blankenship AB. Listed versus unlisted numbers in telephone-survey samples. *Journal of Advertising Research* 1977; **17**:39-42.

5. Blot WJ, McLaughlin JK, Winn DM, Austin DF, Greenberg RS, Preston-Martin S, Bernstein L, Schoenberg JB, Stemhagen A, Fraumeni JF. Smoking and drinking in relation to oral and pharyngeal cancer. *Cancer Research* 1988; **48**:3282-7.

6. Brick JM, Waksberg J. Avoiding sequential sampling with random digit dialing. *Survey Methodology* 1991; **17**:27-41.

7. Brogan, D. *Design and Implementation of the Area Control Sample Survey of the NCI WISH (Women's Interview Study of Health) Study*. Technical Report, Emory University, Atlanta, GA, 1991. (1991a)

8. Brogan, D. *Instructions for Mapping, Counting, Listing, Sampling, and Enumeration Procedures in the WISH Study*. Supervisory Manual, Emory University, Atlanta, GA, 1991. (1991b)

9. Brogan, D. *Women's Interview Study of Health (WISH) Interviewer's Manual for Screening and Enumerating; Area Control Sample Manual.* Emory University, Atlanta, GA, 1991. (1991c)

10. Brunner JA, Brunner GA. Are voluntarily unlisted telephone subscribers really different? *Journal of Marketing Research* 1971; **8**:121-4.

11. Burkheimer GJ, Levinsohn JR. Implementing the Mitofsky-Waksberg sampling design with accelerated sequential replacement. *In*: Groves RM, *et al.*, eds. *Telephone Survey Methodology*. New York, NY: John Wiley & Sons, Inc., 1988.

12. Bureau of the Census. *Census of Population and Housing, 1990: Public Use Microdata Sample U.S. Technical Documentation.* Washington: The Bureau, 1992.

13. Casady RJ, Lepkowski JM. Optimal allocation for stratified telephone survey designs. *Proceedings of the American Statistical Association, Section on Survey Research Methods* 1991:111-16.

14. Converse JM. *Survey Research in the United States: Roots and Emergence 1890-1960.* Berkeley and Los Angeles, CA: University of California Press, 1987.

15. Corey CR, Freeman HF. Use of telephone interviewing in health care research. *Health Services Research* 1990; **25**:129-44.

16. Cummings KM. Random digit dialing: a sampling technique for telephone surveys. *Public Opinion Quarterly* 1979; **43**:233-44.

17. Glasser GJ, Metzger GD. Random digit dialing as a method of telephone sampling. *Journal of Marketing Research* 1972; **9**:59-64.

18. Glasser GJ, Metzger GD. National estimates of nonlisted telephone households and their characteristics. *Journal of Marketing Research* 1975; **12**:359-61.

19. Greenberg ER. Random digit dialing for control selection: a review and a caution on its use in studies of childhood cancer. *American Journal of Epidemiology* 1990; **131**:1-5.

20. Groves RM. *Survey Errors and Survey Costs*. New York, NY: John Wiley & Sons, Inc., 1989.

21. Groves RM, Kahn RL. *Surveys by Telephone: A National Comparison with Personal Interviews*. New York, NY: Academic Press, 1979.

22. Groves RM, Lyberg LE. An overview of nonresponse issues in telephone surveys. *In*: Groves RM, *et al*., eds. *Telephone Survey Methodology*. New York, NY: John Wiley & Sons, Inc., 1988.

23. Harlow BL, Hartge P. Telephone household screening and interviewing. *American Journal of Epidemiology* 1983; **117**:632-3.

24. Hartge P, Brinton LA, Rosenthal JF, Cahill JI, Hoover RN, Waksberg J. Random digit dialing in selecting a population-based control group. *American Journal of Epidemiology* 1984; **120**:825-33.

25. Hansen MH, Hurwitz WN, Madow WG. *Sample Survey Methods and Theory. Volume I: Methods and Applications*. New York, NY: John Wiley & Sons, Inc., 1953.

26. Hiatt RA, Bawol RD. Alcoholic beverage consumption and breast cancer incidence. *American Journal of Epidemiology* 1984; **120**: 676-83.

27. Kelsey JL, Thompson WD, Evans AS. *Methods in Observational Epidemiology*. New York, NY: Oxford University Press, 1986.

28. Kish L. *Survey Sampling*. New York, NY: John Wiley & Sons, Inc., 1965.

29. Lepkowski JM. Telephone sampling methods in the United States. *In*: Groves RM, *et al*., eds. *Telephone Survey Methodology*. New York, NY: John Wiley & Sons, Inc., 1988.

30. Longnecker MP, Berlin JA, Orza MJ, *et al*. A meta-analysis of alcohol consumption in relation to risk of breast cancer. *JAMA* 1988; **260**: 652-6.

31. Longnecker MP. Re: "The evaluation of the data collection process for a multicenter, population-based case-control design" (letter). *American Journal of Epidemiology* 1989; **129**:1311.

32. Mohadjer L. Stratification of prefix areas for sampling random populations. *In*: Groves RM, *et al*., eds. *Telephone Survey Methodology*. New York, NY: John Wiley & Sons, Inc., 1988.

33. O'Connell DL, Hulka BS, Chambless LE, *et al.* Cigarette smoking, alcohol consumption, and breast cancer risk. *JNCI* 1987; **78**: 229-34.

34. Olsen GW, Mandel JS. Selection of elderly controls using random digit dialing. *American Journal of Public Health* 1988; **78**:1487-8.

35. Olson SH, Kelsey JL, Pearson TA, Levin B. Evaluation of random digit dialing as a method of control selection in case-control studies. *American Journal of Epidemiology* 1992; **135**:210-22.

36. Orden SR, Dyer AR, Liu K, Perkins L, Ruth KJ, Burke G, Manolio TA. Random digit dialing in Chicago CARDIA: comparison of individuals with unlisted and listed telephone numbers. *American Journal of Epidemiology* 1992; **135**:697-709.

37. Perneger TV, Myers TL, Klag MJ, Whelton PK. Effectiveness of the Waksberg telephone sampling method for the selection of population controls. *American Journal of Epidemiology* 1993; **138**:574-84.

38. Potter FJ, McNeill JJ, Williams SR, Whitman MA. List-assisted RDD telephone surveys. *Proceedings of the American Statistical Association, Section on Survey research Methods* 1991:117-22.

39. Potthoff RF. Some generalizations of the Mitofsky-Waksberg technique for random digit dialing. *JASA* 1987; **82**: 409-18.

40. Rea LM, Parker RA. *Designing and Conducting Survey Research: A Comprehensive Guide.* San Francisco, CA: Jossey-Bass Publishers, 1992.

41. Rothman KJ. *Modern Epidemiology.* Boston, MA: Little, Brown and Company, 1986.

42. Schatzkin A, Jones DY, Hoover RN, *et al.* Alcohol consumption and breast cancer in the epidemiologic follow-up study of the first National Health and Nutrition Examination Survey. *N Engl J Med* 1987; **316**: 1169-73.

43. Schlesselman JJ. *Case-control Studies: Design, Conduct and Analysis.* New York, NY: Oxford University Press, 1982.

44. Shah BV, Barnwell BG, Bieler GS. *SUDAAN User's Manual, Release 6.40.* Research Triangle Park, NC: Research Triangle Institute, 1995.

45. Sudman S. *Applied Sampling*. New York, NY: Academic Press, 1976.

46. Thornberg OT, Massey JT. Trends in United States Telephone Coverage Across Time and Subgroups. *In*: Groves RM, *et al.*, eds. *Telephone Survey Methodology*. New York, NY: John Wiley & Sons, Inc., 1988.

47. Tucker C, Lepkowski JM, Casady RJ, Groves RM. Commercial residential telephone lists: their characteristics and uses in survey design. *Social Science Computer Review* 1992; **10**: 158-72.

48. Waksberg J. Sampling methods for random digit dialing. *Journal of the American Statistical Association* 1978; **73**:40-6.

49. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. *American Journal of Epidemiology* 1992; **135**:1029-41.

50. Warwick DP, Lininger CA. *The Sample Survey: Theory and Practice.* New York, NY: McGraw-Hill Book Company, 1975.

51. Willett WC, Stampfer MJ, Colditz GA, *et al.* Moderate alcohol consumption and the risk of breast cancer. *N Engl J Med.* 1987; **316**: 1174-80.

52. Wingo PA, Ory HW, Layde PM, Lee NC, *et al.* The evaluation of the data collection process for a multicenter, population-based case-control design. *American Journal of Epidemiology* 1988; **128**:206-17.

7. OTHER ACCOMPLISHMENTS AND OUTCOMES

As noted in Section 1.2 on the specific aims of this Army grant, one major objective was for me to refocus part of my career activity in breast cancer research while maintaining my long standing interest in sample survey methodology. Below are activities during the time of the Army grant directed toward this specific aim.

I presented three papers based on the RDD/area research reported above and am scheduled to present a seminar on these results in October, 1997 at NCI/NIH. In addition, I will present these results at the "An Era of Hope" meeting in Washington, DC in November, 1997. See BIBLIOGRAPHY (Section 9) for a list of these presentations.

One of my master's (MSPH) students, Maxine Denniston, carried out, under my direction, the analyses which used Census data and reported these results in her master's thesis in 1997. See BIBLIOGRAPHY (Section 9) for the citation.

I am first author on two manuscripts in preparation, based on the RDD/area comparisons reported above. See BIBLIOGRAPHY (Section 9) for the citations.

As part of my Army grant I spent a sabbatical year at NCI/NIH. While there I collaborated with several epidemiologists on manuscripts based on the WISH study of risk factors for breast cancer in younger women, and I continue this activity even now. This experience trained me in the analytical approaches typically used by breast cancer epidemiologists, and I was able to make statistically based contributions to several manuscripts. To date, I am co-author on eight published or in-press WISH publications, co-author on two submitted WISH manuscripts, and co-author on two WISH manuscripts now in the final stages of preparation. These manuscripts are listed in the BIBLIOGRAPHY (Section 9).

During my last month at NCI I submitted a grant application in response to an NCI RFA on breast cancer and was awarded in late 1995 a grant "Portion of Breast Cancer Due to Known/Suspected Factors". This project is still in process.

Since my return to Emory in fall of 1995 I have become involved in three projects related to breast cancer or cancer in general. First, I have a contract with the American Cancer Society national headquarters office in Atlanta to serve as a statistical and sample survey collaborator in the review and redesign of the National Cancer DataBase (NCDB); I am working primarily with Dr. Phyllis Wingo and her colleagues. Second, I am a sample survey and statistical collaborator on the national Cancer Survivor Study being designed and conducted by the American Cancer Society in Atlanta; I am working primarily with Dr. Frank Baker and his colleagues. Breast cancer is one of the major cancers included in the Cancer Survivor Study. Third, I worked on a small contract with the CDC National Program of Cancer Registries (NPCR) to review the audit sampling program being used to determine the completeness and accuracy of data reported to statewide cancer registries. I worked with Ms. Carol White and her CDC colleagues. These three projects are of high interest to me since they combine my background in sample survey techniques with the refocus of my research interest into breast cancer. My value as a collaborator in these projects is directly related to the activities carried out under the Army grant, i.e. my year at NCI collaborating with epidemiologists there and my comparison research on area and RDD sampling.

The activities conducted under the Army grant have further advanced my research ideas in sample survey design and analysis. Based on this I have written a few manuscripts and have several in draft form; see the BIBLIOGRAPHY (Section 9).

8. APPENDICES--TABLES

See following pages for Tables 1 through 12.

Table 1 HOUSING UNIT ENUMERATION OUTCOMES -- AREA and RDD SAMPLES

OUTCOME	NUMBER OF HOUSING UNITS with OUTCOME AREA Sampling	RDD Sampling
TOTAL HUs/TELEPHONE NUMBERS	3804	12033
TOTAL NONCHARGEABLES	486	6542
Not a Housing Unit -Area	25	
Vacant -Area	461	
Non residential number -RDD		2176
Out of 3 county area -RDD		408
Non-working number -RDD		3563
No answer, nonresidential -RDD		395
TOTAL NONALLOCATED -RDD		49
No answer, working number		40
No answer, no information		9
TOTAL UNSUCCESSFUL	168	515
No one ever at home -Area	59	
Unavailable	47	
No answer, residential number -RDD		112
Language Problem	6	27
Refused	56	304
Maximum Contact -RDD		72
TOTAL ENUMERATIONS	3150	4927
Complete Enumerations	3150	4572
No eligible women	1097	2033
Eligible -RDD		2539
Eligible, 0 selected -Area	1266	
Eligible, 1 selected - Area	777	
Eligible, 2 selected - Area	10	
Partial * Enumerations -RDD		355
TOTAL CHARGEABLE HUs/HHs	3318	5464**
PERCENT OF CHARGEABLE HUs/HHs		
Refused	1.7%	5.6%
Language Problems	0.2%	0.5%
Contact Problems	3.2%	3.4%
Successful Enumerations***	94.9%	90.2% (RDD1) 83.7% (RDD2)
PERCENTAGE OF ENUMERATED HUs/HHs		
With no women aged 20-54	34.8%	41.3%

* Partial enumeration for RDD means it was determined that the household contained women aged 20-54, and the number and ages of such women, but that further contact information (names, address) was not obtained.

** The total chargeable households is 5464.25 with a percentage of "nonallocated" households counted as chargeable.

*** Two different enumeration rates (percent of successful enumerations) can be calculated for the RDD sample: partial enumerations considered as successful (method RDD1) or as unsuccessful (method RDD2). See text pages 33-35 for detailed explanation.

Table 2 INTERVIEW OUTCOMES -- AREA and RDD SAMPLES

OUTCOME	NUMBER OF SELECTED WOMEN with OUTCOME AREA SAMPLE	RDD SAMPLE
SELECTED for Interview	802	898
EXCLUDED from Interview	8	80
Wrong gender	0	2
Out of age range	5	14
Residence out of 3 county area	0	5
Duplicate	2	6
Partial Complete, not converted*	0	42
Other	1	11
ELIGIBLE for Interview	794	818
INTERVIEW Outcome		
Completed interview	640	652
Refused interview	105	113
Deceased	0	3
Unavailable	15	13
Language Problem	9	9
Too ill	2	3
Moved from 3 county area	22	23
Other	1	2
INTERVIEW Response Rate**	80.6%	75.8% (RDD1) 79.7% (RDD2)
OVERALL Survey Response Rate** (enumeration rate x interview rate)	76.5%	68.4% (RDD1) 66.7% (RDD2)

* Partial complete, not converted means that when enumeration was performed, it was determined that the household contained women aged 20-54, but information necessary for contacting those women if they were selected for interview (names and/or addresses) was not obtained. On 42 women selected for interview from households that gave only partial enumeration information, further attempts via telephone to obtain name and/or address information on the selected women were unsuccessful. They were closed out and no further attempts were made to interview them.

** Two different interview response rates and their corresponding overall survey response rates can be calculated for the RDD sample depending on whether the partial enumerations were considered as successful (method RDD1) or unsuccessful (method RDD2) for the calculation of the enumeration response rate. See text pages 36-37 for a more detailed explanation.

Table 3 Estimated Race Distribution (with standard errors) by Sample Type and County
Women Aged 30-54 Years, Metro Atlanta 1990-1992
Census, Area, and RDD Samples -- Weighted/Clustered Analysis

COUNTY / SAMPLE TYPE / RACE GROUP	Sample Size	Weighted Size	Row Percent	SE Row Percent
All 3 Counties	Chi-sq = 3.58*	df = 2	p = 0.17	
Census Sample				
Total	14086	325994	100.00	0.00
Non-blacks	9317	212161	65.08	0.44
RDD Sample				
Total	640	259385	100.00	0.00
Non-blacks	470	184522	71.14	3.19
Area Sample				
Total	626	333260	100.00	0.00
Non-blacks	406	221546	66.48	5.31
Cobb County	Chi-sq = 8.73	df = 2	p = 0.01	
Census Sample				
Total	3920	94103	100.00	0.00
Non-blacks	3609	86021	91.41	0.50
RDD Sample				
Total	174	75431	100.00	0.00
Non-blacks	169	72947	96.71	1.54
Area Sample				
Total	191	85705	100.00	0.00
Non-blacks	172	78013	91.03	3.96
DeKalb County	Chi-sq = 0.12	df = 2	p = 0.94	
Census Sample				
Total	4844	108796	100.00	0.00
Non-blacks	2890	62333	57.29	0.78
RDD Sample				
Total	220	89916	100.00	0.00
Non-blacks	146	52978	58.92	5.89
Area Sample				
Total	206	107311	100.00	0.00
Non-blacks	111	59072	55.05	10.77
Fulton County	Chi-sq = 4.21	df = 2	p = 0.12	
Census Sample				
Total	5322	123095	100.00	0.00
Non-blacks	2818	63807	51.84	0.74
RDD Sample				
Total	246	94039	100.00	0.00
Non-blacks	155	58596	62.31	5.21
Area Sample				
Total	229	140244	100.00	0.00
Non-blacks	123	84461	60.22	9.66

* Chi-sq for independence of sample type and race group (percentages)

Table 4 Estimated Percentage of Women Born in the US (with standard errors) by Sample Type and County
Women Aged 30-54 Years, Metro Atlanta 1990-1992
Census, Area, and RDD Samples -- Weighted/Clustered Analysis

COUNTY SAMPLE TYPE BIRTHPLACE	Sample Size	Weighted Size	Row Percent	SE Row Percent
All 3 Counties	Chi-sq = 7.81*		df = 2	p = 0.02
Census				
Total	14086	325994	100.00	0.00
Born in US	13179	304837	93.51	0.22
RDD Sample				
Total	640	259385	100.00	0.00
Born in US	605	245519	94.65	1.12
Area Sample				
Total	626	333260	100.00	0.00
Born in US	594	320406	96.14	0.85
Cobb County	Chi-sq = 7.09		df = 2	p = 0.03
Census				
Total	3920	94103	100.00	0.00
Born in US	3661	88064	93.58	0.41
RDD Sample				
Total	174	75431	100.00	0.00
Born in US	166	72695	96.37	1.39
Area Sample				
Total	191	85705	100.00	0.00
Born in US	180	82579	96.35	1.16
DeKalb County	Chi-sq = 5.84		df = 2	p = 0.054
Census				
Total	4844	108796	100.00	0.00
Born in US	4449	99629	91.57	0.44
RDD Sample				
Total	220	89916	100.00	0.00
Born in US	204	84523	94.00	1.59
Area Sample				
Total	206	107311	100.00	0.00
Born in US	193	102612	95.62	1.40
Fulton County	Chi-sq = 0.78		df = 2	p = 0.68
Census				
Total	5322	123095	100.00	0.00
Born in US	5069	117144	95.17	0.33
RDD Sample				
Total	246	94039	100.00	0.00
Born in US	235	88300	93.90	2.45
Area Sample				
Total	229	140244	100.00	0.00
Born in US	221	135214	96.41	1.57

* Chi-sq for independence of sample type and birthplace (percentages)

Table 5 Estimated Percentage of High School Graduates (with standard errors) by Sample Type and County
Women Aged 30-54 Years, Metro Atlanta 1990-1992
Census, Area, and RDD Samples -- Weighted/Clustered Analysis

COUNTY SAMPLE TYPE GRADUATE HS	Sample Size	Weighted Size	Row Percent	SE Row Percent
All 3 Counties	Chi-sq = 18.90*		df = 2	p < 0.01
Census Sample				
Total	14086	325994	100.00	0.00
YES	12413	287244	88.11	0.29
RDD Sample				
Total	640	259385	100.00	0.00
YES	572	241873	93.25	1.14
Area Sample				
Total	626	333260	100.00	0.00
YES	556	303406	91.04	1.76
Cobb County	Chi-sq = 7.62		df = 2	p = 0.02
Census Sample				
Total	3920	94103	100.00	0.00
YES	3575	85836	91.21	0.48
RDD Sample				
Total	174	75431	100.00	0.00
YES	159	71909	95.33	1.35
Area Sample				
Total	191	85705	100.00	0.00
YES	178	80838	94.32	2.72
DeKalb County	Chi-sq = 9.36		df = 2	p = 0.01
Census Sample				
Total	4844	108796	100.00	0.00
YES	4343	97015	89.17	0.50
RDD Sample				
Total	220	89916	100.00	0.00
YES	204	85038	94.58	1.55
Area Sample				
Total	206	107311	100.00	0.00
YES	183	99249	92.49	2.39
Fulton County	Chi-sq = 4.98		df = 2	p = 0.08
Census Sample				
Total	5322	123095	100.00	0.00
YES	4495	104393	84.81	0.52
RDD Sample				
Total	246	94039	100.00	0.00
YES	209	84926	90.31	2.38
Area Sample				
Total	229	140244	100.00	0.00
YES	195	123318	87.93	3.48

* Chi-sq for independence of sample type and High School Graduation (percentages)

Table 6 Estimated Percentage of High School Graduates (with standard errors) by Sample Type and Race
Women Aged 30-54 Years, Metro Atlanta 1990-1992
Census, Area, and RDD Samples -- Weighted/Clustered Analysis

RACE GROUP SAMPLE TYPE GRADUATE HS	Sample Size	Weighted Size	Row Percent	SE Row Percent
All Races	Chi-sq = 18.90*		df = 2	p < 0.01
Census Sample				
Total	14086	325994	100.00	0.00
YES	12413	287244	88.11	0.29
RDD Sample				
Total	640	259385	100.00	0.00
YES	572	241873	93.25	1.14
Area Sample				
Total	626	333260	100.00	0.00
YES	556	303406	91.04	1.76
Blacks	Chi-sq = 8.06		df = 2	p = 0.02
Census Sample				
Total	4769	113833	100.00	0.00
YES	3773	90958	79.90	0.62
RDD Sample				
Total	170	74864	100.00	0.00
YES	135	66029	88.20	2.82
Area Sample				
Total	220	111714	100.00	0.00
YES	174	93722	83.90	3.08
Non-blacks	Chi-sq = 7.54		df = 2	p = 0.02
Census Sample				
Total	9317	212161	100.00	0.00
YES	8640	196286	92.52	0.30
RDD Sample				
Total	470	184522	100.00	0.00
YES	437	175844	95.30	0.97
Area Sample				
Total	406	221546	100.00	0.00
YES	382	209683	94.65	1.96

* Chi-sq for independence of sample type and High School Graduation (percentages)

Table 7 Estimated Distribution of Marital Status (with standard errors) by Sample Type and County
Women Aged 30-54 Years, Metro Atlanta 1990-1992
Census, Area, and RDD Samples -- Weighted/Clustered Analysis

COUNTY / SAMPLE TYPE / Marital Status	Sample Size	Weighted Size	Row Percent	SE Row Percent
All 3 Counties	Chi-sq = 13.42*		df = 4	p = 0.01
Census Sample				
Total	14086	325994	100.00	0.00
Married	8509	194650	59.71	0.44
W/D/S**	3005	70999	21.78	0.38
Separated	2572	60345	18.51	0.35
RDD Sample				
Total	640	259385	100.00	0.00
Married	428	176061	67.88	2.75
W/D/S	97	39510	15.23	2.11
Separated	115	43814	16.89	2.07
Area Sample				
Total	626	333260	100.00	0.00
Married	405	224151	67.26	4.31
W/D/S	94	51874	15.57	3.15
Separated	127	57235	17.17	2.41
Cobb County	Chi-sq = 10.37		df = 4	p = 0.03
Census Sample				
Total	3920	94103	100.00	0.00
Married	2861	67378	71.60	0.77
W/D/S	417	10895	11.58	0.57
Separated	642	15830	16.82	0.63
RDD Sample				
Total	174	75431	100.00	0.00
Married	131	62074	82.29	3.49
W/D/S	12	4106	5.44	1.76
Separated	31	9251	12.26	3.06
Area Sample				
Total	191	85705	100.00	0.00
Married	141	62703	73.16	5.64
W/D/S	15	8812	10.28	4.11
Separated	35	14190	16.56	3.26
DeKalb County	Chi-sq = 3.35		df = 4	p = 0.50
Census Sample				
Total	4844	108796	100.00	0.00
Married	2847	63235	58.12	0.77
W/D/S	1094	25162	23.13	0.67
Separated	903	20399	18.75	0.60
RDD Sample				
Total	220	89916	100.00	0.00
Married	139	55777	62.03	4.54
W/D/S	43	20348	22.63	4.54
Separated	38	13791	15.34	2.82
Area Sample				
Total	206	107311	100.00	0.00
Married	136	74212	69.16	6.99
W/D/S	33	16634	15.50	4.78
Separated	37	16465	15.34	3.90
Fulton County	Chi-sq = 13.34		df = 4	p = 0.01
Census Sample				
Total	5322	123095	100.00	0.00
Married	2801	64037	52.02	0.72
W/D/S	1494	34942	28.39	0.66
Separated	1027	24116	19.59	0.57
RDD Sample				
Total	246	94039	100.00	0.00
Married	158	58210	61.90	4.87
W/D/S	42	15056	16.01	3.28
Separated	46	20772	22.09	4.28
Area Sample				
Total	229	140244	100.00	0.00
Married	128	87237	62.20	8.31
W/D/S	46	26428	18.84	6.11
Separated	55	26580	18.95	4.52

* Chi-sq for independence of sample type and marital status (percentages)

** Widowed, divorced, or single (never married)

Table 8 Estimated Distribution of Marital Status (with standard errors) by Sample Type and Race
Women Aged 30-54 Years, Metro Atlanta 1990-1992
Census, Area, and RDD Samples--Weighted/Clustered Analysis

RACE GROUP / SAMPLE TYPE / Marital Status	Sample Size	Weighted Size	Row Percent	SE Row Percent
All Races	Chi-sq = 13.42*		df = 4	p = 0.01
Census Sample				
Total	14086	325994	100.00	0.00
Married	8509	194650	59.71	0.44
W/D/S**	3005	70999	21.78	0.38
Separated	2572	60345	18.51	0.35
RDD Sample				
Total	640	259385	100.00	0.00
Married	428	176061	67.88	2.75
W/D/S	97	39510	15.23	2.11
Separated	115	43814	16.89	2.07
Area Sample				
Total	626	333260	100.00	0.00
Married	405	224151	67.26	4.31
W/D/S	94	51874	15.57	3.15
Separated	127	57235	17.17	2.41
Blacks	Chi-sq = 3.27		df = 4	p = 0.51
Census Sample				
Total	4769	113833	100.00	0.00
Married	2039	49112	43.14	0.76
W/D/S	1722	40544	35.62	0.73
Separated	1008	24177	21.24	0.62
RDD Sample				
Total	170	74864	100.00	0.00
Married	81	30749	41.07	4.48
W/D/S	53	23293	31.11	4.88
Separated	36	20822	27.81	4.75
Area Sample				
Total	220	111714	100.00	0.00
Married	99	48870	43.75	5.48
W/D/S	63	33993	30.43	6.34
Separated	58	28851	25.83	3.84
Non-blacks	Chi-sq = 15.79		df = 4	p < 0.01
Census Sample				
Total	9317	212161	100.00	0.00
Married	6470	145538	68.60	0.52
W/D/S	1283	30455	14.35	0.40
Separated	1564	36168	17.05	0.41
RDD Sample				
Total	470	184522	100.00	0.00
Married	347	145312	78.75	2.71
W/D/S	44	16218	8.79	2.00
Separated	79	22991	12.46	1.90
Area Sample				
Total	406	221546	100.00	0.00
Married	306	175282	79.12	4.14
W/D/S	31	17881	8.07	2.45
Separated	69	28384	12.81	2.63

* Chi-sq for independence of sample type and marital status (percentages)
** Widowed, divorced, or single (never married)

Table 9 Estimated Telephone Coverage* (with standard errors) by Sample Type and County
Women Aged 30-54 Years, Metro Atlanta 1990-1992
Census and Area Samples -- Weighted/Clustered Analysis

COUNTY / SAMPLE TYPE / Residential Telephone	Sample Size	Weighted Size	Row Percent	SE Row Percent
All 3 Counties	Chi-sq = 0.01**		df = 1	p-value = 0.92
Census Sample				
Total	14086	325994	100.00	0.00
YES	13696	317408	97.37	0.14
Area Sample				
Total	624	332584	100.00	0.00
YES	613	324225	97.49	1.14
Cobb County***				
Census Sample				
Total	3920	94103	100.00	0.00
YES	3888	93341	99.19	0.15
Area Sample				
Total	190	85282	100.00	0.00
YES	190	85282	100.00	0.00
DeKalb County	Chi-sq = 0.16		df = 1	p-value = 0.69
Census Sample				
Total	4844	108796	100.00	0.00
YES	4744	106639	98.02	0.21
Area Sample				
Total	206	107311	100.00	0.00
YES	202	104438	97.32	1.76
Fulton County	Chi-sq = 0.08		df = 1	p-value = 0.77
Census Sample				
Total	5322	123095	100.00	0.00
YES	5064	117428	95.40	0.30
Area Sample				
Total	228	139992	100.00	0.00
YES	221	134504	96.08	2.35

* Defined as percentage of women living in a household with a residential telephone
** Chi square for independence of sample type and residential telephone coverage (percentages)
*** Chi square not calculated for Cobb county because point estimate for area sample is 100%

Table 10 Estimated Telephone Coverage* (with standard errors) by Sample Type and Race
Women Aged 30-54 Years, Metro Atlanta 1990-1992
Census and Area Samples -- Weighted/Clustered Analysis

RACE GROUP SAMPLE TYPE Residential Telephone	Sample Size	Weighted Size	Row Percent	SE Row Percent
All Races	Chi-sq = 0.01**	df = 1	p-value = 0.92	
Census Sample				
Total	14086	325994	100.00	0.00
YES	13696	317408	97.37	0.14
Area Sample				
Total	624	332584	100.00	0.00
YES	613	324225	97.49	1.14
Non-Blacks	Chi-sq = 0.01	df = 1	p-value = 0.90	
Census Sample				
Total	9317	212161	100.00	0.00
YES	9232	210464	99.20	0.09
Area Sample				
Total	404	220871	100.00	0.00
YES	402	219297	99.29	0.72
Blacks	Chi-sq = 0.01	df = 1	p-value = 0.99	
Census Sample				
Total	4769	113833	100.00	0.00
YES	4464	106944	93.95	0.36
Area Sample				
Total	220	111714	100.00	0.00
YES	211	104928	93.93	2.90

* Defined as percentage of women living in a household with a residential telephone
** Chi square for independence of sample type and residential telephone coverage (percentages)

Table 11 Age Distribution (with standard errors) Sample Type and Race
Women Aged 30-54 Years, Area and RDD Samples, Metro Atlanta 1990-1992
Unweighted/Unclustered and Unweighted/Clustered Analyses

RACE GROUP SAMPLE TYPE AGE GROUP	Sample Size	Row Percent	SE Row Percent unclustered	SE Row Percent clustered
All Races	Chi-sq = 12.60*	df = 3	p-value = 0.01 (unclustered)	
	Chi-sq = 12.00	df = 3	p-value = 0.01 (clustered)	
Area Sample				
Total	626	100.00	0.00	0.00
30-39 yrs	129	20.61	1.62	1.92
40-44 yrs	166	26.52	1.77	1.75
45-49 yrs	193	30.83	1.85	2.07
50-54 yrs	138	22.04	1.66	1.66
RDD Sample				
Total	640	100.00	0.00	0.00
30-39 yrs	118	18.44	1.53	1.38
40-44 yrs	153	23.91	1.69	1.62
45-49 yrs	172	26.88	1.75	1.75
50-54 yrs	197	30.78	1.83	1.76
Non-blacks	Chi-sq = 6.49	df = 3	p-value = 0.09 (unclustered)	
	Chi-sq = 5.94	df = 3	p-value = 0.12 (clustered)	
Area Sample				
Total	406	100.00	0.00	0.00
30-39 yrs	74	18.23	1.92	2.17
40-44 yrs	93	22.91	2.09	2.21
45-49 yrs	137	33.74	2.35	2.67
50-54 yrs	102	25.12	2.15	2.20
RDD Sample				
Total	470	100.00	0.00	0.00
30-39 yrs	80	17.02	1.73	1.62
40-44 yrs	104	22.13	1.92	1.88
45-49 yrs	133	28.30	2.08	2.09
50-54 yrs	153	32.55	2.16	2.12
Blacks	Chi-sq = 5.00	df = 3	p-value = 0.17 (unclustered)	
	Chi-sq = 4.60	df = 3	p-value = 0.20 (clustered)	
Area Sample				
Total	220	100.00	0.00	0.00
30-39 yrs	55	25.00	2.92	3.30
40-44 yrs	73	33.18	3.18	2.90
45-49 yrs	56	25.45	2.94	2.68
50-54 yrs	36	16.36	2.50	2.43
RDD Sample				
Total	170	100.00	0.00	0.00
30-39 yrs	38	22.35	3.20	2.95
40-44 yrs	49	28.82	3.48	2.96
45-49 yrs	39	22.94	3.23	3.45
50-54 yrs	44	25.88	3.36	3.34

* Chi square for independence of sample type and age group (percentages)

Table 12 Race Distribution (with standard errors) by Sample Type and County
Women Aged 30-54 Years, Area and RDD Samples
Metro Atlanta 1990-1992
Unweighted/Unclustered and Unweighted/Clustered Analysis

COUNTY SAMPLE TYPE RACE GROUP	Sample Size	Row Percent	SE Row Percent unclustered	SE Row Percent clustered
All 3 Counties	Chi-sq = 10.99*		df = 1	p < 0.01 (unclustered)
	Chi-sq = 2.94		df = 1	p = 0.09 (clustered)
Area Sample				
Total	626	100.00	0.00	0.00
Non-blacks	406	64.86	1.91	4.28
RDD Sample				
Total	640	100.00	0.00	0.00
Non-blacks	470	73.44	1.75	2.55
Cobb County	Chi-sq = 7.47		df = 1	p = 0.01 (unclustered)
	Chi-sq = 2.40		df = 1	p = 0.12 (clustered)
Area Sample				
Total	191	100.00	0.00	0.00
Non-blacks	172	90.05	2.17	3.90
RDD Sample				
Total	174	100.00	0.00	0.00
Non-blacks	169	97.13	1.27	1.24
DeKalb County	Chi-sq = 6.71		df = 1	p = 0.01 (unclustered)
	Chi-sq = 1.77		df = 1	p = 0.18 (clustered)
Area Sample				
Total	206	100.00	0.00	0.00
Non-blacks	111	53.88	3.48	8.05
RDD Sample				
Total	220	100.00	0.00	0.00
Non-blacks	146	66.36	3.19	4.42
Fulton County	Chi-sq = 4.15		df = 1	p = 0.04 (unclustered)
	Chi-sq = 1.15		df = 1	p = 0.28 (clustered)
Area Sample				
Total	229	100.00	0.00	0.00
Non-blacks	123	53.71	3.30	7.29
RDD Sample				
Total	246	100.00	0.00	0.00
Non-blacks	155	63.01	3.08	4.34

* Chi-square for independence of sample type and race group (percentages)

9. BIBLIOGRAPHY OF PRODUCTS RELATED TO ARMY GRANT

Presentations on Comparison of RDD and Area Sampling

1. **Brogan, Donna**. "Area Probability Sampling vs. Random Digit Dialing (RDD) Sampling: An Empirical Comparison". Invited seminar presented by **Donna Brogan** on 4/19/96 to the Statistics Department at Purdue University, West Lafayette, IN. This seminar was in conjunction with Dr. Brogan's receipt of the Purdue University Distinguished Alumna Award (from the School of Arts and Sciences).

2. **Brogan, Donna**. "Area Probability Sampling vs. Random Digit Dialing (RDD) Sampling: An Empirical Comparison". Invited seminar presented by **Donna Brogan** on 10/25/96 at the Mathematics and Computer Science Department, Georgia State University, Atlanta.

3. **Brogan, Donna** and Maxine Denniston. "Population-Based Sampling for Case-Control Studies". Contributed paper presented by **Donna Brogan** at annual meeting of International Biometric Society (Eastern North American Region—ENAR) on 3/24/97 in Memphis, TN.

4. **Brogan, Donna**. "Population-Based Sampling for Breast Cancer Case-Control Studies". Poster session to be presented by **Donna Brogan** at "An Era of Hope" meeting sponsored by the U.S. Army's Breast Cancer Research Program, Oct. 31 - Nov. 4, 1997, Washington, DC.

5. **Brogan, Donna**. "Comparison of RDD Sampling and Area Probability Sampling in a Breast Cancer Case-Control Study". Invited seminar to be presented by **Donna Brogan** at NCI/NIH, Rockville, MD on October 30, 1997.

Publications on Comparison of RDD and Area Sampling

6. Denniston, Maxine (1997). <u>A Comparison of Area and Random Digit Dialing Sampling</u> . MSPH thesis. Rollins School of Public Health, Emory University, Atlanta, GA. Advisor: **Donna Brogan**. Readers: Jonathan Liff and Elaine Flagg (faculty in Epidemiology Dept.). Ms. Denniston was a finalist for the Shephard Award for an outstanding MSPH thesis and presented her thesis at the award competition on 5/9/97.

Manuscripts in Preparation on Comparison of RDD and Area Sampling

7. **Brogan, Donna,** Maxine Denniston, (other probable authors are Jonathan Liff, Louise Brinton, Ralph Coates,). "RDD versus Area Sampling for Population-Based Controls in a Breast Cancer Case-Control Study".

8. **Brogan, Donna**, Maxine Denniston (other probable authors are Jonathan Liff, Ralph Coates, Louise Brinton). "Empirical Comparison of Area and RDD Sampling".

Published Papers Based on the WISH Breast Cancer Case-Control Study

9. Weiss, Helen, L. Brinton, **Donna Brogan**, R. Coates, M. Gammon, K. Malone, J. Schoenberg, C. Swanson (1996). "Epidemiology of In Situ and Invasive Breast Cancer in Women Aged Under 45". British Journal of Cancer, 73, 1298-1305.

10. Weiss, Helen, L Brinton, N Potischman, **Donna Brogan**, R Coates, M Gammon, K Malone, J Schoenberg (1997). "Prenatal and Perinatal Risk Factors for Breast Cancer in Young Women", Epidemiology, 8(2), 181-187.

11. Swanson, Christine, R Coates, K Malone, M Gammon, J Schoenberg, **Donna Brogan**, M McAdams, N Potischman, R Hoover, L Brinton (in press). "Alcohol Consumption and Breast Cancer Risk Among Women Under Age 45", Epidemiology.

12. Cook, Linda, J. Daling, L. Voigt, M. DeHart, K. Malone, J. Stanford, N. Weiss, L. Brinton, M. Gammon, **Donna Brogan** (1997). "Characteristics of Women with and without Breast Augmentation", J. Amer. Medical Assoc., 277(20), May 28, 1612-1617.

13. Potischman, Nancy, C Swanson, R Coates, H Weiss, **Donna Brogan**, J Stanford, J Schoenberg, M Gammon, L. Brinton (in press). "Dietary Relationships with Early Onset (Age <45) Breast Cancer in a Case-Control Study: Influence of Chemotherapy Treatment", Cancer Causes and Control .

14. Weiss, Helen, L Brinton, N Potischman, **Donna Brogan**, R Coates, M Gammon, K Malone, J Schoenberg (in press). "Breast Cancer Risk in Young Women and History of Selected Medical Conditions", Int. J. Epidemiology.

15. Gammon, Marilie, J. Schoenberg, J. Britton, J. Kelsey, R. Coates, **Donna Brogan**, N. Potischman, C. Swanson, J. Daling, J. Stanford, L. Brinton (in press). "Recreational Physical Activity and Breast Cancer Risk among Women under Age 45 Years". Submitted to American Journal of Epidemiology.

16. Brinton, Louise, J Benichou, M Gammon, **Donna Brogan**, R Coates, J Schoenberg (in press). "Ethnicity and Variation in Breast Cancer Incidence", International Journal of Cancer.

Submitted Manuscripts Based on the WISH Breast Cancer Case-Control Study

17. Potischman, Nancy, H Weiss, L Brinton, R Coates, J Daling, **Donna Brogan**, J Schoenberg. "Adolescent Diet and Risk of Breast Cancer among Young Women", submitted to Journal of National Cancer Institute, June, 1997.

18. Troisi, Rebecca, H Weiss, R Hoover, N Potischman, C Swanson, **Donna Brogan**, R Coates, M Gammon, K Malone, L Brinton. "Pregnancy Characteristics and Maternal Risk of Breast Cancer", submitted to Epidemiology, June, 1997.

Manuscripts in Preparation Based on the WISH Breast Cancer Case-Control Study

19. Troisi, Rebecca, H Weiss, M Rossing, L Brinton et al (with **Donna Brogan**). "Fertility Problems and Breast Cancer Risk in Young Women: A Case-Control Study".

20. Gammon, Marilie, J Schoenberg, J Britton, J Kelsey, J Stanford, K Malone, R Coates, **Donna Brogan**, N Potischman, C Swanson, L Brinton. "Electric Blanket Use and Breast Cancer Risk among Younger Women".

Published Manuscripts Related to Sample Survey Research of the U.S. Army Grant

21. Malilay, Josephine, D Flanders and **Donna Brogan** (1996). "A Modified Cluster Sampling Method for Post-Disaster Rapid Needs Assessment", Bulletin of the World Health Organization, 74(4), 399-405.

22. **Brogan, Donna** (in press, to be published 1998). "Pitfalls of Using Standard Statistical Software Packages for Sample Survey Data", **invited chapter** in Encyclopedia of Biostatistics, edited by Peter Armitage and Theodore Colton, John Wiley, New York.

Manuscripts in Preparation Related to Research Done on the U.S. Army Grant

23. **Brogan, Donna**, Dana Flanders, Jeanne Calle, and others. "Estimation of Breast Cancer Mortality Attributable Risk in the Cancer Prevention Study-II".

24. Lamar, Verna and **Donna Brogan**. "Breast Cancer Screening Practices: Behavioral Risk Factor Surveillance System, 1992".

25. **Brogan, Donna**, Maxine Denniston, Erica Frank. "Comparison of SAS vs. SUDAAN for Analysis of a Stratified Random Sample of Women Physicians".

26. **Brogan, Donna**, Danni Daniels, Debbie Rolka, Fred Marsteller. "Comparison of SAS vs. SUDAAN for Analysis of a Stratified Random Sample of the Youthful Offender Population".

27. **Brogan, Donna**, Danni Daniels, Fred Marsteller. "Combining Two Random Digit Dialing Surveys for Decreased Cost".

28. Elon, Lisa, **Donna Brogan**, Melissa Adams, Carol Hogue. "Nonresponse in the PRAMS Statewide Surveys".

10. PERSONNEL PAID FROM ARMY GRANT

P.I. Donna Brogan, Ph.D.
Research Assistant: Maxine Denniston, MSPH

Most of the money from the Army grant paid a portion of Dr. Brogan's salary over a period of two years to conduct the research reported on herein.